Probabilidad y estadística

para ingeniería y ciencias

NOVENA EDICIÓN

Walpole • Myers • Myers

ALWAYS LEARNING

PEARSON

Probabilidad y estadística para ingeniería y ciencias

Probabilidad y estadística para ingeniería y ciencias

Novena edición

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

Traducción
Leticia Esther Pineda Ayala
Traductora especialista en estadística

Revisión técnica
Roberto Hernández Ramírez
Departamento de Física y Matemáticas
División de Ingeniería y Tecnologías
Universidad de Monterrey

Linda Margarita Medina Herrera
Departamento de Física y Matemáticas
Escuela de Diseño, Ingeniería y Arquitectura
Instituto Tecnológico y de Estudios Superiores de Monterrey,
Campus Ciudad de México

PEARSON

Datos de catalogación bibliográfica

RONALD E. WALPOLE, RAYMOND H. MYERS,
SHARON L. MYERS Y KEYING YE

Probabilidad y estadística para ingeniería y ciencias
Novena edición

PEARSON EDUCACIÓN, México, 2012
ISBN: 978-607-32-1417-9
Área: Ingeniería

Formato: 18.5 × 23.5 cm Páginas: 816

Authorized translation from the English language edition, entitled *PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS 9th E*dition, by *RONALD E. WALPOLE, RAYMOND H. MYERS, SHARON L. MYERS and KEYING YE,* published by Pearson Education, Inc., publishing as Pearson, Copyright © 2012. All rights reserved.
ISBN 9780321629111

Traducción autorizada de la edición en idioma inglés, titulada *PROBABILIDAD Y ESTADÍSTICA PARA INGENIERÍA Y CIENCIAS 9ª* edición por *RONALD E. WALPOLE, RAYMOND H. MYERS, SHARON L. MYERS y KEYING YE,* publicada por Pearson Education, Inc., publicada como Pearson, Copyright © 2012. Todos los derechos reservados.

Esta edición en español es la única autorizada.

Edición en español
Dirección Educación Superior: Mario Contreras
Editor sponsor: Gabriela López Ballesteros
e-mail: gabriela.lopezballesteros@pearson.com
Editor de desarrollo: Felipe Hernández Carrasco
Supervisor de Producción: Juan José García Guzmán
Diseño de portada: Dream Studio/Edgar Maldonado
Gerencia editorial
Educación Superior Latinoamérica: Marisa de Anta

NOVENA EDICIÓN, 2012

D.R. © 2012 por Pearson Educación de México, S.A. de C.V.
Atlacomulco 500-5o. piso
Col. Industrial Atoto
53519, Naucalpan de Juárez, Estado de México

Cámara Nacional de la Industria Editorial Mexicana. Reg. núm. 1031.

Reservados todos los derechos. Ni la totalidad ni parte de esta publicación pueden reproducirse, registrarse o transmitirse, por un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea electrónico, mecánico, fotoquímico, magnético o electroóptico, por fotocopia, grabación o cualquier otro, sin permiso previo por escrito del editor.

El préstamo, alquiler o cualquier otra forma de cesión de uso de este ejemplar requerirá también la autorización del editor o de sus representantes.

ISBN VERSIÓN IMPRESA: 978-607-32-1417-9
ISBN VERSIÓN E-BOOK: 978-607-32-1418-6
ISBN E-CHAPTER: 978-607-32-1419-3

Impreso en México. *Printed in Mexico*.
1 2 3 4 5 6 7 8 9 0 - 15 14 13 12

PEARSON

www.pearsonenespañol.com

AGRADECIMIENTOS

Pearson agradece a los profesores usuarios de esta obra y a los centros de estudio por su apoyo y retroalimentación, elementos fundamentales para esta nueva edición de Probabilidad y estadística para ingeniería y ciencias.

COLOMBIA

Escuela Colombiana de Ingeniería
Departamento de Matemáticas
Susana Rondón Troncoso

Pontificia Universidad Javeriana Cali
Departamento de Ciencias Naturales y Matemáticas
Daniel Enrique González Gómez
María del Pilar Marín Gaviria
Sandra Milena Ramírez Buelvas

Universidad Católica de Colombia
Departamento de Ciencias Básicas
Queeny Madueño Pinto

Universidad de La Salle
Departamento de Ciencias Básicas
Maribel Méndez Cortés
Martha Tatiana Jiménez Valderrama
Milton Armando Reyes Villamil
Myrian Elena Vergara Morales

COSTA RICA

Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Producción Industrial
Ivannia Hasbum Fernández

Universidad de Costa Rica
Escuela de Estadística
Facultad de Ciencias Económicas
Ana Teresa Garita Salas

MÉXICO

Estado de México

Facultad de Estudios Superiores Cuautitlán C-4
Armando Aguilar Márquez
Fermín Cervantes Martínez
Héctor Coss Garduño
Juan Carlos Axotla García
Miguel de Nazareth Pineda Becerril
Vicente Vázquez Juárez

Tecnológico de Estudios Superiores de Coacalco
María de la Luz Dávila Flores
Martha Nieto López
Héctor Feliciano Martínez Osorio
Jeanette López Alanís

Tecnológico de Estudios Superiores de Ecatepec
Héctor Rodríguez Carmona
Ángel Hernández Estrada
Daniel Jaimes Serrano
Ramón Jordán Rocha

Jalisco

Universidad de Guadalajara
Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI)
Departamento de Matemáticas
Agustín Rodríguez Martínez
Carlos Florentino Melgoza Cañedo
Cecilia Garibay López
Dalmiro García Nava
Deliazar Pantoja Espinoza
Gloria Arroyo Cervantes
Javier Nava Gómez
Jorge Luis Rodríguez Gutiérrez
José Ángel Partida Ibarra
José de Jesús Bernal Casillas
José de Jesús Cabrera Chavarría
José de Jesús Rivera Prado
José Solís Rodríguez
Julieta Carrasco García
Laura Esther Cortés Navarro
Lizbeth Díaz Caldera
Maribel Sierra Fuentes
Mario Alberto Prado Alonso
Osvaldo Camacho Castillo
Rosalía Buenrostro Arceo
Samuel Rosalío Cuevas

Universidad del Valle de México, Zapopan
Departamento de Ingeniería
Abel Vázquez Pérez
Irene Isabel Navarro González
Jorge Eduardo Aguilar Rosas
Miguel Arturo Barreiro González

Sinaloa

Instituto Tecnológico de Culiacán
Ciencias Básicas
Cecilia Norzagaray Gámez

Instituto Tecnológico de Los Mochis
Ciencias Básicas
Jesús Alberto Báez Torres

Contenido

Prefacio ..... **xv**

1 Introducción a la estadística y al análisis de datos **1**

1.1 Panorama general: inferencia estadística, muestras, poblaciones y el papel de la probabilidad 1

1.2 Procedimientos de muestreo; recolección de los datos 7

1.3 Medidas de localización: la media y la mediana de una muestra 11

Ejercicios 13

1.4 Medidas de variabilidad 14

Ejercicios 17

1.5 Datos discretos y continuos 17

1.6 Modelado estadístico, inspección científica y diagnósticos gráficos 18

1.7 Tipos generales de estudios estadísticos: diseño experimental, estudio observacional y estudio retrospectivo 27

Ejercicios 30

2 Probabilidad **35**

2.1 Espacio muestral 35

2.2 Eventos 38

Ejercicios 42

2.3 Conteo de puntos muestrales 44

Ejercicios 51

2.4 Probabilidad de un evento 52

2.5 Reglas aditivas 56

Ejercicios 59

2.6 Probabilidad condicional, independencia y regla del producto 62

Ejercicios 69

2.7 Regla de Bayes 72

Ejercicios 76

Ejercicios de repaso 77

2.8 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 79

3 Variables aleatorias y distribuciones de probabilidad 81

3.1 Concepto de variable aleatoria 81
3.2 Distribuciones discretas de probabilidad 84
3.3 Distribuciones de probabilidad continua 87
Ejercicios 91
3.4 Distribuciones de probabilidad conjunta 94
Ejercicios 104
Ejercicios de repaso 107
3.5 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 109

4 Esperanza matemática 111

4.1 Media de una variable aleatoria 111
Ejercicios 117
4.2 Varianza y covarianza de variables aleatorias 119
Ejercicios 127
4.3 Medias y varianzas de combinaciones lineales de variables aleatorias 128
4.4 Teorema de Chebyshev 135
Ejercicios 137
Ejercicios de repaso 139
4.5 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 142

5 Algunas distribuciones de probabilidad discreta 143

5.1 Introducción y motivación 143
5.2 Distribuciones binomial y multinomial 143
Ejercicios 150
5.3 Distribución hipergeométrica 152
Ejercicios 157
5.4 Distribuciones binomial negativa y geométrica 158
5.5 Distribución de Poisson y proceso de Poisson 161
Ejercicios 164
Ejercicios de repaso 166
5.6 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 169

6 Algunas distribuciones continuas de probabilidad 171

6.1 Distribución uniforme continua 171
6.2 Distribución normal 172
6.3 Áreas bajo la curva normal 176
6.4 Aplicaciones de la distribución normal 182
Ejercicios 185
6.5 Aproximación normal a la binomial 187
Ejercicios 193
6.6 Distribución gamma y distribución exponencial 194
6.7 Distribución chi cuadrada 200
6.8 Distribución beta 201
6.9 Distribución logarítmica normal 201
6.10 Distribución de Weibull (opcional) 203
Ejercicios 206
Ejercicios de repaso 207
6.11 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 209

7 Funciones de variables aleatorias (opcional) 211

7.1 Introducción 211
7.2 Transformaciones de variables 211
7.3 Momentos y funciones generadoras de momentos 218
Ejercicios 222

8 Distribuciones de muestreo fundamentales y descripciones de datos 225

8.1 Muestreo aleatorio 225
8.2 Algunos estadísticos importantes 227
Ejercicios 230
8.3 Distribuciones muestrales 232
8.4 Distribución muestral de medias y el teorema del límite central 233
Ejercicios 241
8.5 Distribución muestral de S^2 243
8.6 Distribución t 246
8.7 Distribución F 251
8.8 Gráficas de cuantiles y de probabilidad 254
Ejercicios 259
Ejercicios de repaso 260

8.9 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 262

9 Problemas de estimación de una y dos muestras 265

9.1 Introducción 265
9.2 Inferencia estadística 265
9.3 Métodos de estimación clásicos 266
9.4 Una sola muestra: estimación de la media 269
9.5 Error estándar de una estimación puntual 276
9.6 Intervalos de predicción 277
9.7 Límites de tolerancia 280
Ejercicios 282
9.8 Dos muestras: estimación de la diferencia entre dos medias 285
9.9 Observaciones pareadas 291
Ejercicios 294
9.10 Una sola muestra: estimación de una proporción 296
9.11 Dos muestras: estimación de la diferencia entre dos proporciones 300
Ejercicios 302
9.12 Una sola muestra: estimación de la varianza 303
9.13 Dos muestras: estimación de la proporción de dos varianzas 305
Ejercicios 307
9.14 Estimación de la máxima verosimilitud (opcional) 307
Ejercicios 312
Ejercicios de repaso 313
9.15 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 316

10 Pruebas de hipótesis de una y dos muestras 319

10.1 Hipótesis estadísticas: conceptos generales 319
10.2 Prueba de una hipótesis estadística 321
10.3 Uso de valores *P* para la toma de decisiones en la prueba de hipótesis 331
Ejercicios 334
10.4 Una sola muestra: pruebas respecto a una sola media 336
10.5 Dos muestras: pruebas sobre dos medias 342
10.6 Elección del tamaño de la muestra para la prueba de medias 349
10.7 Métodos gráficos para comparar medias 354
Ejercicios 356
10.8 Una muestra: prueba sobre una sola proporción 361
10.9 Dos muestras: pruebas sobre dos proporciones 363
Ejercicios 365
10.10 Pruebas de una y dos muestras referentes a varianzas 366
Ejercicios 369

10.11 Prueba de la bondad de ajuste....371
10.12 Prueba de independencia (datos categóricos)....374
10.13 Prueba de homogeneidad....376
10.14 Estudio de caso de dos muestras....380
Ejercicios....382
Ejercicios de repaso....384
10.15 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos....387

11 Regresión lineal simple y correlación....389
11.1 Introducción a la regresión lineal....389
11.2 El modelo de regresión lineal simple (RLS)....390
11.3 Mínimos cuadrados y el modelo ajustado....394
Ejercicios....398
11.4 Propiedades de los estimadores de mínimos cuadrados....400
11.5 Inferencias sobre los coeficientes de regresión....403
11.6 Predicción....408
Ejercicios....411
11.7 Selección de un modelo de regresión....414
11.8 El método del análisis de varianza....414
11.9 Prueba para la linealidad de la regresión: datos con observaciones repetidas....416
Ejercicios....421
11.10 Gráficas de datos y transformaciones....424
11.11 Estudio de caso de regresión lineal simple....428
11.12 Correlación....430
Ejercicios....435
Ejercicios de repaso....436
11.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos....442

12 Regresión lineal múltiple y ciertos modelos de regresión no lineal....443
12.1 Introducción....443
12.2 Estimación de los coeficientes....444
12.3 Modelo de regresión lineal en el que se utilizan matrices....447
Ejercicios....450
12.4 Propiedades de los estimadores de mínimos cuadrados....453
12.5 Inferencias en la regresión lineal múltiple....455
Ejercicios....461
12.6 Selección de un modelo ajustado mediante la prueba de hipótesis....462

12.7 Caso especial de ortogonalidad (opcional) 467
Ejercicios....471
12.8 Variables categóricas o indicadoras 472
Ejercicios....476
12.9 Métodos secuenciales para la selección del modelo....476
12.10 Estudio de los residuales y violación de las suposiciones (verificación del modelo) 482
12.11 Validación cruzada, C_p, y otros criterios para la selección del modelo 487
Ejercicios....494
12.12 Modelos especiales no lineales para condiciones no ideales....496
Ejercicios....500
Ejercicios de repaso 501
12.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 506

13 Experimentos con un solo factor: generales 507
13.1 Técnica del análisis de varianza....507
13.2 La estrategia del diseño de experimentos 508
13.3 Análisis de varianza de un factor: diseño completamente aleatorizado (ANOVA de un factor)....509
13.4 Pruebas de la igualdad de varias varianzas....516
Ejercicios....518
13.5 Comparaciones de un grado de libertad....520
13.6 Comparaciones múltiples....523
Ejercicios....529
13.7 Comparación de un conjunto de tratamientos en bloques 532
13.8 Diseños de bloques completos aleatorizados....533
13.9 Métodos gráficos y verificación del modelo....540
13.10 Transformaciones de datos en el análisis de varianza 543
Ejercicios....545
13.11 Modelos de efectos aleatorios....547
13.12 Estudio de caso 551
Ejercicios....553
Ejercicios de repaso 555
13.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 559

14 Experimentos factoriales (dos o más factores) 561
14.1 Introducción....561
14.2 Interacción en el experimento de dos factores....562
14.3 Análisis de varianza de dos factores 565
Ejercicios....575

14.4 Experimentos de tres factores 579
Ejercicios 586
14.5 Experimentos factoriales para efectos aleatorios y modelos mixtos 588
Ejercicios 592
Ejercicios de repaso 594
14.6 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 596

15 Experimentos factoriales 2^k y fracciones 597
15.1 Introducción 597
15.2 El factorial 2^k: cálculo de efectos y análisis de varianza 598
15.3 Experimento factorial 2^k sin réplicas 604
Ejercicios 609
15.4 Experimentos factoriales en un ajuste de regresión 612
15.5 El diseño ortogonal 617
Ejercicios 625
15.6 Experimentos factoriales fraccionarios 626
15.7 Análisis de experimentos factoriales fraccionados 632
Ejercicios 634
15.8 Diseños de fracciones superiores y de filtrado 636
15.9 Construcción de diseños de resolución III y IV, con 8, 16 y 32 puntos de diseño 637
15.10 Otros diseños de resolución III de dos niveles; los diseños de Plackett-Burman 638
15.11 Introducción a la metodología de superficie de respuesta 639
15.12 Diseño robusto de parámetros 643
Ejercicios 652
Ejercicios de repaso 653
15.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 654

16 Estadística no paramétrica 655
16.1 Pruebas no paramétricas 655
16.2 Prueba de rango con signo 660
Ejercicios 663
16.3 Prueba de la suma de rangos de Wilcoxon 665
16.4 Prueba de Kruskal-Wallis 668
Ejercicios 670
16.5 Pruebas de rachas 671
16.6 Límites de tolerancia 674

16.7 Coeficiente de correlación de rango 674
Ejercicios....677
Ejercicios de repaso 679

17 Control estadístico de la calidad 681
17.1 Introducción....681
17.2 Naturaleza de los límites de control....683
17.3 Objetivos de la gráfica de control 683
17.4 Gráficas de control para variables....684
17.5 Gráficas de control para atributos 697
17.6 Gráficas de control de cusum....705
Ejercicios de repaso 706

18 Estadística bayesiana 709
18.1 Conceptos bayesianos....709
18.2 Inferencias bayesianas 710
18.3 Estimados bayesianos mediante la teoría de decisión 717
Ejercicios....718

Bibliografía 721

Apéndice A: Tablas y demostraciones estadísticas....725

Apéndice B: Respuestas a los ejercicios impares (no de repaso) 769

Índice....785

Prefacio

Enfoque general y nivel matemático

Al elaborar la novena edición, nuestro interés principal no fue tan sólo agregar material nuevo sino brindar claridad y mejor comprensión. Este objetivo se logró en parte al incluir material nuevo al final de los capítulos, lo cual permite que se relacionen mejor. Con cierto afecto llamamos "contratiempos" a los comentarios que aparecen al final de los capítulos, pues son muy útiles para que los estudiantes recuerden la idea general y la forma en que cada capítulo se ajusta a esa imagen; así como para que entiendan las limitaciones y los problemas que resultarían por el uso inadecuado de los procedimientos. Los proyectos para la clase favorecen una mayor comprensión de cómo se utiliza la estadística en el mundo real, por lo que añadimos algunos proyectos en varios capítulos. Tales proyectos brindan a los estudiantes la oportunidad de trabajar solos o en equipo, y de reunir sus propios datos experimentales para realizar inferencias. En algunos casos, el trabajo implica un problema cuya solución ejemplifica el significado de un concepto, o bien, favorece la comprensión empírica de un resultado estadístico importante. Se ampliaron algunos de los ejemplos anteriores y se introdujeron algunos nuevos para crear "estudios de caso", los cuales incluyen un comentario para aclarar al estudiante un concepto estadístico en el contexto de una situación práctica.

En esta edición seguimos haciendo énfasis en el equilibrio entre la teoría y las aplicaciones. Utilizamos el cálculo y otros tipos de conceptos matemáticos, por ejemplo, de álgebra lineal, casi al mismo nivel que en ediciones anteriores. Las herramientas analíticas para la estadística se cubren de mejor manera utilizando el cálculo en los casos donde el análisis se centra en las reglas de los conceptos de probabilidad. En los capítulos 2 a 10 se destacan las distribuciones de probabilidad y la inferencia estadística. En los capítulos 11 a 15, en los cuales se estudian la regresión lineal y el análisis de varianza, se aplica un poco de álgebra lineal y matrices. Los estudiantes que utilizan este libro deben haber cursado el equivalente a un semestre de cálculo diferencial e integral. El álgebra lineal es útil aunque no indispensable, siempre y cuando el instructor no cubra la sección sobre regresión lineal múltiple del capítulo 12 utilizando álgebra de matrices. Al igual que en las ediciones anteriores, y con la finalidad de desafiar al estudiante, muchos ejercicios se refieren a aplicaciones científicas y de ingeniería a la vida real. Todos los conjuntos de datos asociados con los ejercicios están disponibles para descargar del sitio web http://www.pearsonenespañol.com/walpole.

Resumen de los cambios en la novena edición

- Para brindar una mayor comprensión del uso de la estadística en el mundo real, en varios capítulos se agregaron proyectos para la clase. Los estudiantes tienen que generar o reunir sus propios datos experimentales y realizar inferencias a partir de ellos.
- Se agregaron más estudios de caso y otros se ampliaron para ayudar a los usuarios a comprender los métodos estadísticos que se presentan en el contexto de una situación real. Por ejemplo, la interpretación de los límites de confianza, los límites de predicción y los límites de tolerancia se exponen utilizando situaciones de la vida real.
- Se agregaron "contratiempos" al final de algunos capítulos y en otros se ampliaron los que ya se incluían. El objetivo de dichos comentarios es presentar cada capítulo en el contexto de la idea general y analizar la forma en que los capítulos se relacionan entre sí. Otro objetivo es advertir acerca del uso inadecuado de las técnicas estadísticas examinadas en el capítulo.
- El capítulo 1 se mejoró y ahora incluye más estadísticos de una sola cifra y técnicas gráficas. También se incluyó nuevo material fundamental sobre muestreo y diseño experimental.
- Los ejemplos que se agregaron en el capítulo 8 sobre las distribuciones de muestreo tienen la finalidad de motivar a los estudiantes a realizar las pruebas de hipótesis y de los valores *P*. Esto los prepara para el material más avanzado sobre los temas que se presentan en el capítulo 10.
- El capítulo 12 contiene más información sobre el efecto que tiene una sola variable de regresión en un modelo que presenta una gran colinealidad con otras variables.
- El capítulo 15 ahora introduce material sobre el importante tema de la metodología de superficie de respuesta (MSR). El uso de las variables del ruido en la MSR permite ejemplificar los modelos de la media y la varianza (superficie de respuesta doble).
- En el capítulo 15 se introduce el diseño compuesto central.
- El capítulo 18 incluye más ejemplos y un mejor análisis de cómo se utilizan los métodos bayesianos para la toma de decisiones estadísticas.

Contenido y planeación del curso

Este libro está diseñado para un curso de uno o dos semestres. Un plan razonable para el curso de un semestre podría incluir los capítulos 1 a 10, lo cual daría como resultado un programa que concluye con los fundamentos de la estimación y la prueba de hipótesis. Los profesores que desean que los estudiantes aprendan la regresión lineal simple podrían incluir una parte del capítulo 11. Para quienes deseen incluir el análisis de varianza en vez de la regresión, el curso de un semestre podría incluir el capítulo 13 en vez de los capítulos 11 y 12. El capítulo 13 trata el tema del análisis de varianza de un factor. Otra opción consiste en eliminar partes de los capítulos 5 o 6, así como el capítulo 7. Al hacer esto se omitirían las distribuciones discretas o continuas, mismas que incluyen la binomial negativa, la geométrica, la gamma, la de Weibull, la beta y la logarítmica normal. Otros contenidos que se podrían omitir en un programa de un semestre son la estimación de máxima verosimilitud, la predicción y los límites de tolerancia del

capítulo 9. El programa para un semestre suele ser flexible, dependiendo del interés que el profesor tenga en la regresión, el análisis de varianza, el diseño experimental y los métodos de superficie de respuesta (capítulo 15). Existen varias distribuciones discretas y continuas (capítulos 5 y 6) que tienen aplicaciones en diversas áreas de la ingeniería y las ciencias.

Los capítulos 11 a 18 incluyen una gran cantidad de material que se podría agregar al segundo semestre, en caso de que se eligiera un curso de dos semestres. El material sobre la regresión lineal simple y múltiple se estudia en los capítulos 11 y 12, respectivamente. El capítulo 12 puede ser muy flexible. La regresión lineal múltiple incluye "temas especiales", como variables categóricas o indicadoras, métodos secuenciales para la selección de modelos, por ejemplo, la regresión por etapas, el estudio de residuales para la detección de violaciones de supuestos, la validación cruzada y el uso de los estadísticos PRESS, así como el de C_p y la regresión logística. Se hace hincapié en el uso de regresores ortogonales, un precursor del diseño experimental en el capítulo 15. Los capítulos 13 y 14 ofrecen hasta cierto grado material abundante sobre el análisis de varianza (ANOVA), con modelos fijos, aleatorios y mixtos. En el capítulo 15 se destaca la aplicación de los diseños con dos niveles en el contexto de los experimentos factoriales fraccionarios y completos (2^k). También se ejemplifican los diseños especiales de selección. En el capítulo 15 se incluye asimismo una nueva sección sobre la metodología de superficie de respuesta (MSR), para ejemplificar el uso del diseño experimental con la finalidad de encontrar condiciones óptimas de proceso. Se analiza el ajuste de un modelo de segundo orden utilizando un diseño complejo central. La MSR se amplía para abarcar el análisis de problemas sobre el diseño de un parámetro robusto. Las variables de ruido se utilizan para ajustar modelos dobles de superficie de respuesta. Los capítulos 16, 17 y 18 incluyen una cantidad moderada de material sobre estadística no paramétrica, control de calidad e inferencia bayesiana.

El capítulo 1 es un bosquejo de la inferencia estadística, presentada a un nivel matemático sencillo, pero de manera más amplia que en la octava edición con el propósito de examinar más detalladamente los estadísticos de una sola cifra y las técnicas gráficas. Este capítulo está diseñado para brindar a los estudiantes una presentación preliminar de los conceptos fundamentales que les permitirán entender los detalles posteriores de mayor complejidad. Se presentan conceptos clave sobre muestreo, recolección de datos y diseño experimental, así como los aspectos rudimentarios de las herramientas gráficas y la información que se obtiene a partir de un conjunto de datos. También se agregaron las gráficas de tallo y hojas, y las de caja y bigotes. Las gráficas están mejor organizadas y etiquetadas. El análisis de la incertidumbre y la variación en un sistema se ilustra de forma detallada. Se incluyen ejemplos de cómo clasificar las características importantes de un sistema o proceso científico, y esas ideas se ilustran en ambientes prácticos, como procesos de manufactura, estudios biomédicos, y estudios de sistemas biológicos y científicos de otros tipos. Se efectúa una comparación entre el uso de los datos discretos y continuos; también se hace un mayor énfasis en el uso de modelos y de la información con respecto a los modelos estadísticos que se logran obtener mediante las herramientas gráficas.

En los capítulos 2, 3 y 4 se estudian los conceptos básicos de probabilidad, así como las variables aleatorias discretas y continuas. Los capítulos 5 y 6 se enfocan en las distribuciones discretas y continuas específicas, así como en las relaciones que existen entre ellas. En estos capítulos también se destacan ejemplos de aplicaciones de las distribuciones en estudios reales científicos y de ingeniería. Los estudios de caso, los ejemplos y una gran cantidad de ejercicios permiten a los estudiantes practicar el uso de tales distribuciones. Los proyectos permiten la aplicación práctica de estas distribuciones en la vida

real mediante el trabajo en equipo. El capítulo 7 es el más teórico del libro; en él se expone la transformación de variables aleatorias, y podría ser que no se utilice a menos que el instructor desee impartir un curso relativamente teórico. El capítulo 8 contiene material gráfico, el cual amplía el conjunto básico de herramientas gráficas presentadas y ejemplificadas en el capítulo 1. Aquí se analizan las gráficas de probabilidad y se ilustran con ejemplos. El muy importante concepto de las distribuciones de muestreo se presenta de forma detallada, y se proporcionan ejemplos que incluyen el teorema del límite central y la distribución de una varianza muestral en una situación de muestreo independiente y normal. También se presentan las distribuciones t y F para motivar a los estudiantes a utilizarlas en los capítulos posteriores. El nuevo material del capítulo 8 ayuda a los estudiantes a conocer la importancia de la prueba de hipótesis mediante la presentación del concepto del valor P.

El capítulo 9 contiene material sobre la estimación puntual y de intervalos de una muestra y dos muestras. Un análisis detallado y con ejemplos destaca las diferencias entre los tipos de intervalos (intervalos de confianza, intervalos de predicción e intervalos de tolerancia). Un estudio de caso ilustra los tres tipos de intervalos estadísticos en el contexto de una situación de manufactura. Este estudio de caso destaca las diferencias entre los intervalos, sus fuentes y los supuestos en que se basan, así como cuáles son los intervalos que requieren diferentes tipos de estudios o preguntas. Se añadió un método de aproximación para las inferencias sobre una proporción. El capítulo 10 inicia con una presentación básica sobre el significado práctico de la prueba de hipótesis, con un énfasis en conceptos fundamentales como la hipótesis nula y la alternativa, el papel que desempeñan la probabilidad y el valor P, así como la potencia de una prueba. Después, se presentan ejemplos de pruebas sobre una o dos muestras en condiciones estándar. También se describe la prueba t de dos muestras con observaciones en pares (apareadas). Un estudio de caso ayuda a los estudiantes a entender el verdadero significado de una interacción de factores, así como los problemas que en ocasiones surgen cuando existen interacciones entre tratamientos y unidades experimentales. Al final del capítulo 10 se incluye una sección muy importante que relaciona los capítulos 9 y 10 (estimación y prueba de hipótesis) con los capítulos 11 a 16, donde se destaca el modelamiento estadístico. Es importante que el estudiante esté consciente de la fuerte relación entre los capítulos mencionados.

Los capítulos 11 y 12 incluyen material sobre la regresión lineal simple y múltiple, respectivamente. En esta edición ponemos mucho más atención en el efecto que tiene la colinealidad entre las variables de regresión. Se presenta una situación que muestra cómo el papel que desempeña una sola variable de regresión depende en gran parte de cuáles son los regresores que la acompañan en el modelo. Después se revisan los procedimientos secuenciales para la selección del modelo (hacia adelante, hacia atrás, por etapas, etcétera) con respecto a este concepto, así como los fundamentos para utilizar ciertos tipos de valores P con tales procedimientos. En el capítulo 12 se estudia material sobre los modelos no lineales con una presentación especial de la regresión logística, la cual tiene aplicaciones en ingeniería y en las ciencias biológicas. El material sobre la regresión múltiple es muy extenso, de manera que, como antes se expuso, plantea una gran flexibilidad. Al final del capítulo 12 se incluye un comentario que lo relaciona con los capítulos 14 y 15. Se agregaron varios elementos para fomentar la comprensión del material en general. Por ejemplo, al final del capítulo se describen algunas dificultades y problemas que podrían surgir. Se indica que existen tipos de respuestas que ocurren de forma natural en la práctica, por ejemplo, respuestas de proporciones, de conteo y muchas otras, con las cuales no se debe utilizar la regresión estándar de mínimos cuadrados

debido a que los supuestos de normalidad no se cumplen, y transgredirlos causaría errores muy graves. Se sugiere utilizar la transformación de datos para reducir el problema en algunos casos. Nuevamente, los capítulos 13 y 14 sobre el tema del análisis de varianza tienen cierta flexibilidad. En el capítulo 13 se estudia el ANOVA de un factor en el contexto de un diseño completamente aleatorio. Algunos temas complementarios incluyen las pruebas sobre las varianzas y las comparaciones múltiples. Se destacan las comparaciones de tratamientos en bloque, junto con el tema de los bloques completos aleatorizados. Los métodos gráficos se extendieron al ANOVA para ayudar al estudiante a complementar la inferencia formal con una inferencia pictórica que facilita la presentación del material a los científicos y a los ingenieros. Se incluye un nuevo proyecto donde los estudiantes incorporan la aleatoriedad adecuada a cada plan, y se utilizan técnicas gráficas y valores *P* en el informe de los resultados. En el capítulo 14 se amplía el material del capítulo 13 para ajustar dos o más factores dentro de una estructura factorial. La presentación del ANOVA en el capítulo 14 incluye la creación de modelos aleatorios y de efectos fijos. En el capítulo 15 se estudia material relacionado con los diseños factoriales 2^k; los ejemplos y los estudios de caso plantean el uso de diseños de selección y fracciones especiales de orden superior del factorial 2^k. Dos elementos nuevos y especiales son la metodología de superficie de respuesta (MSR) y el diseño de parámetros robustos. Son temas que se relacionan en un estudio de caso que describe e ilustra un diseño doble de superficie de respuesta, así como un análisis que incluye el uso de superficies de respuesta de la media y la varianza de procesos.

Programa de cómputo

Los estudios de caso, que inician en el capítulo 8, muestran impresiones de listas de resultados por computadora y material gráfico generado con los programas SAS y MINITAB. El hecho de incluir los cálculos por computadora refleja nuestra idea de que los estudiantes deben contar con la experiencia de leer e interpretar impresiones de listas de resultados y gráficas por computadora, incluso si el software que se utiliza en el libro no coincide con el que utiliza el profesor. La exposición a más de un tipo de programas aumentaría la experiencia de los estudiantes. No hay razones para creer que el programa utilizado en el curso coincidirá con el que el estudiante tendrá que utilizar en la práctica después de graduarse. Cuando sea pertinente, los ejemplos y los estudios de caso en el libro se complementarán con diversos tipos de gráficas residuales, cuantilares, de probabilidad normal y de otros tipos. Tales gráficas se incluyen especialmente en los capítulos 11 a 15.

Complementos

Manual de soluciones para el instructor. Este recurso contiene respuestas a todos los ejercicios del libro y se puede descargar del Centro de Recursos para Profesor de Pearson.

Diapositivas de PowerPoint® ISBN-10: 0-321-73731-8; ISBN-13: 978-0-321-73731-1. Las diapositivas incluyen la mayoría de las figuras y las tablas del libro; se pueden descargar del Centro de Recursos para el Profesor de Pearson.

Reconocimientos

Estamos en deuda con los colegas que revisaron las anteriores ediciones de este libro y que nos dieron muchas sugerencias útiles para esta edición. Ellos son David Groggel, de *Miami University*; Lance Hemlow, de *Raritan Valley Community College*; Ying Ji, de *University of Texas at San Antonio*; Thomas Kline, de *University of Northern Iowa*; Sheila Lawrence, de *Rutgers University*; Luis Moreno, de *Broome County Community College;* Donald Waldman, de *University of Colorado-Boulder* y Marlene Will, de *Spalding University*. También queremos agradecer a Delray Schulz, de *Millersville University*, Roxane Burrows, de *Hocking College* y Frank Chmely por asegurarse de la exactitud de este libro.

Nos gustaría agradecer a la editorial y a los servicios de producción suministrados por muchas personas de Pearson/Prentice Hall, sobre todo a Deirdre Lynch, la editora en jefe, a Christopher Cummings, el editor de adquisiciones, a Christine O'Brien, la editora de contenido ejecutivo, a Tracy Patruno, la editora de producción y a Sally Lifland, la editora de producción. Apreciamos los comentarios y sugerencias útiles de Gail Magin, la correctora de estilo. También estamos en deuda con el Centro de Asesoría Estadística de Virginia Tech, que fue nuestra fuente de muchos conjuntos reales de datos.

R.H.M.
S.L.M.
K.Y.

CAPÍTULO 1

Introducción a la estadística y al análisis de datos

1.1 Panorama general: inferencia estadística, muestras, poblaciones y el papel de la probabilidad

Desde inicios de la década de los ochenta del siglo pasado y hasta lo que ha transcurrido del siglo XXI la industria estadounidense ha puesto una enorme atención en el *mejoramiento de la calidad*. Se ha dicho y escrito mucho acerca del "milagro industrial" en Japón, que comenzó a mediados del siglo XX. Los japoneses lograron el éxito en donde otras naciones fallaron, a saber, en la creación de un entorno que permita la manufactura de productos de alta calidad. Gran parte del éxito de los japoneses se atribuye al uso de *métodos estadísticos* y del pensamiento estadístico entre el personal gerencial.

Empleo de datos científicos

El uso de métodos estadísticos en la manufactura, el desarrollo de productos alimenticios, el software para computadoras, las fuentes de energía, los productos farmacéuticos y muchas otras áreas implican el acopio de información o **datos científicos**. Por supuesto que la obtención de datos no es algo nuevo, ya que se ha realizado por más de mil años. Los datos se han recabado, resumido, reportado y almacenado para su examen cuidadoso. Sin embargo, hay una diferencia profunda entre el acopio de información científica y la **estadística inferencial**. Esta última ha recibido atención legítima en décadas recientes.

La estadística inferencial generó un número enorme de "herramientas" de los métodos estadísticos que utilizan los profesionales de la estadística. Los métodos estadísticos se diseñan para contribuir al proceso de realizar juicios científicos frente a la **incertidumbre** y a la **variación**. Dentro del proceso de manufactura, la densidad de producto de un material específico no siempre será la misma. De hecho, si un proceso es discontinuo en vez de continuo, la densidad de material no sólo variará entre los lotes que salen de la línea de producción (variación de un lote a otro), sino también dentro de los propios lotes. Los métodos estadísticos se utilizan para analizar datos de procesos como el anterior; el objetivo de esto es tener una mejor orientación respecto de cuáles cambios se deben realizar en el proceso para mejorar su **calidad**. En este proceso la calidad bien podría

definirse en relación con su grado de acercamiento a un valor de densidad meta en armonía con *qué parte de las veces* se cumple este criterio de cercanía. A un ingeniero podría interesarle un instrumento específico que se utilice para medir el monóxido de azufre en estudios sobre la contaminación atmosférica. Si el ingeniero dudara respecto de la eficacia del instrumento, tendría que tomar en cuenta dos **fuentes de variación**. La primera es la variación en los valores del monóxido de azufre que se encuentran en el mismo lugar el mismo día. La segunda es la variación entre los valores observados y la cantidad **real** de monóxido de azufre que haya en el aire en ese momento. Si cualquiera de estas dos fuentes de variación es excesivamente grande (según algún estándar determinado por el ingeniero), quizá se necesite remplazar el instrumento. En un estudio biomédico de un nuevo fármaco que reduce la hipertensión, 85% de los pacientes experimentaron alivio; aunque por lo general se reconoce que el medicamento actual o el "viejo" alivia a 80% de los pacientes que sufren hipertensión crónica. Sin embargo, el nuevo fármaco es más caro de elaborar y podría tener algunos efectos colaterales. ¿Se debería adoptar el nuevo medicamento? Éste es un problema con el que las empresas farmacéuticas, junto con la FDA (Federal Drug Administration), se encuentran a menudo (a veces es mucho más complejo). De nuevo se debe tomar en cuenta las necesidades de variación. El valor del "85%" se basa en cierto número de pacientes seleccionados para el estudio. Tal vez si se repitiera el estudio con nuevos pacientes ¡el número observado de "éxitos" sería de 75%! Se trata de una variación natural de un estudio a otro que se debe tomar en cuenta en el proceso de toma de decisiones. Es evidente que tal variación es importante, ya que la variación de un paciente a otro es endémica al problema.

Variabilidad en los datos científicos

En los problemas analizados anteriormente los métodos estadísticos empleados tienen que ver con la variabilidad y en cada caso la variabilidad que se estudia se encuentra en datos científicos. Si la densidad del producto observada en el proceso fuera siempre la misma y siempre fuera la esperada, no habría necesidad de métodos estadísticos. Si el dispositivo para medir el monóxido de azufre siempre diera el mismo valor y éste fuera exacto (es decir, correcto), no se requeriría análisis estadístico. Si entre un paciente y otro no hubiera variabilidad inherente a la respuesta al medicamento (es decir, si el fármaco siempre causara alivio o nunca aliviara), la vida sería muy sencilla para los científicos de las empresas farmacéuticas y de la FDA, y los estadísticos no serían necesarios en el proceso de toma de decisiones. Los investigadores de la estadística han originado un gran número de métodos analíticos que permiten efectuar análisis de datos obtenidos de sistemas como los descritos anteriormente, lo cual refleja la verdadera naturaleza de la ciencia que conocemos como estadística inferencial, a saber, el uso de técnicas que, al permitirnos obtener conclusiones (o inferencias) sobre el sistema científico, nos permiten ir más allá de sólo reportar datos. Los profesionales de la estadística usan leyes fundamentales de probabilidad e inferencia estadística para sacar conclusiones respecto de los sistemas científicos. La información se colecta en forma de **muestras** o conjuntos de **observaciones**. En el capítulo 2 se introduce el proceso de muestreo, el cual se continúa analizando a lo largo de todo el libro.

Las muestras se reúnen a partir de **poblaciones**, que son conjuntos de todos los individuos o elementos individuales de un tipo específico. A veces una población representa un sistema científico. Por ejemplo, un fabricante de tarjetas para computadora podría desear eliminar defectos. Un proceso de muestreo implicaría recolectar información de 50 tarjetas de computadora tomadas aleatoriamente durante el proceso. En este caso la población

sería representada por todas las tarjetas de computadora producidas por la empresa en un periodo específico. Si se lograra mejorar el proceso de producción de las tarjetas para computadora y se reuniera una segunda muestra de tarjetas, cualquier conclusión que se obtuviera respecto de la efectividad del cambio en el proceso debería extenderse a toda la población de tarjetas para computadora que se produzcan en el "proceso mejorado". En un experimento con fármacos se toma una muestra de pacientes y a cada uno se le administra un medicamento específico para reducir la presión sanguínea. El interés se enfoca en obtener conclusiones sobre la población de quienes sufren hipertensión. A menudo, cuando la planeación ocupa un lugar importante en la agenda, es muy importante el acopio de datos científicos en forma sistemática. En ocasiones la planeación está, por necesidad, bastante limitada. Con frecuencia nos enfocamos en ciertas propiedades o características de los elementos u objetos de la población. Cada característica tiene importancia de ingeniería específica o, digamos, biológica para el "cliente", el científico o el ingeniero que busca aprender algo acerca de la población. Por ejemplo, en uno de los casos anteriores la calidad del proceso se relacionaba con la densidad del producto al salir del proceso. Un(a) ingeniero(a) podría necesitar estudiar el efecto de las condiciones del proceso, la temperatura, la humedad, la cantidad de un ingrediente particular, etcétera. Con ese fin podría mover de manera sistemática estos **factores** a cualesquiera niveles que se sugieran, de acuerdo con cualquier prescripción o **diseño experimental** que se desee. Sin embargo, un científico silvicultor que está interesado en estudiar los factores que influyen en la densidad de la madera en cierta clase de árbol no necesariamente tiene que diseñar un experimento. Este caso quizá requiera un **estudio observacional**, en el cual los datos se acopian en el campo pero no es posible seleccionar de antemano los **niveles de los factores**. Ambos tipos de estudio se prestan a los métodos de la inferencia estadística. En el primero, la calidad de las inferencias dependerá de la planeación adecuada del experimento. En el segundo, el científico está a expensas de lo que pueda recopilar. Por ejemplo, si un agrónomo se interesara en estudiar el efecto de la lluvia sobre la producción de plantas sería lamentable que recopilara los datos durante una sequía.

Es bien conocida la importancia del pensamiento estadístico para los administradores y el uso de la inferencia estadística para el personal científico. Los investigadores obtienen mucho de los datos científicos. Los datos proveen conocimiento acerca del fenómeno científico. Los ingenieros de producto y de procesos aprenden más en sus esfuerzos fuera de línea para mejorar el proceso. También logran una comprensión valiosa al reunir datos de producción (supervisión en línea) sobre una base regular, lo cual les permite determinar las modificaciones que se requiere realizar para mantener el proceso en el nivel de calidad deseado.

En ocasiones un científico sólo desea obtener alguna clase de resumen de un conjunto de datos representados en la muestra. En otras palabras, no requiere estadística inferencial. En cambio, le sería útil un conjunto de estadísticos o la **estadística descriptiva**. Tales números ofrecen un sentido de la ubicación del centro de los datos, de la variabilidad en los datos y de la naturaleza general de la distribución de observaciones en la muestra. Aunque no se incorporen métodos estadísticos específicos que lleven a la **inferencia estadística**, se puede aprender mucho. A veces la estadística descriptiva va acompañada de gráficas. El software estadístico moderno permite el cálculo de **medias**, **medianas**, **desviaciones estándar** y otros estadísticos de una sola cifra, así como el desarrollo de gráficas que presenten una "huella digital" de la naturaleza de la muestra. En las secciones siguientes veremos definiciones e ilustraciones de los estadísticos y descripciones de recursos gráficos como histogramas, diagramas de tallo y hojas, diagramas de dispersión, gráficas de puntos y diagramas de caja.

El papel de la probabilidad

En los capítulos 2 a 6 de este libro se presentan los conceptos fundamentales de la probabilidad. Un estudio concienzudo de las bases de tales conceptos permitirá al lector comprender mejor la inferencia estadística. Sin algo de formalismo en teoría de probabilidad, el estudiante no podría apreciar la verdadera interpretación del análisis de datos a través de los métodos estadísticos modernos. Es muy natural estudiar probabilidad antes de estudiar inferencia estadística. Los elementos de probabilidad nos permiten cuantificar la fortaleza o "confianza" en nuestras conclusiones. En este sentido, los conceptos de probabilidad forman un componente significativo que complementa los métodos estadísticos y ayuda a evaluar la consistencia de la inferencia estadística. Por consiguiente, la disciplina de la probabilidad brinda la transición entre la estadística descriptiva y los métodos inferenciales. Los elementos de la probabilidad permiten expresar la conclusión en el lenguaje que requieren los científicos y los ingenieros. El ejemplo que sigue permite al lector comprender la noción de un valor-*P*, el cual a menudo proporciona el "fundamento" en la interpretación de los resultados a partir del uso de métodos estadísticos.

Ejemplo 1.1: Suponga que un ingeniero se encuentra con datos de un proceso de producción en el cual se muestrean 100 artículos y se obtienen 10 defectuosos. Se espera y se anticipa que ocasionalmente habrá artículos defectuosos. Obviamente estos 100 artículos representan la muestra. Sin embargo, se determina que, a largo plazo, la empresa sólo puede tolerar 5% de artículos defectuosos en el proceso. Ahora bien, los elementos de probabilidad permiten al ingeniero determinar qué tan concluyente es la información muestral respecto de la naturaleza del proceso. En este caso la **población** representa conceptualmente todos los artículos posibles en el proceso. Suponga que averiguamos que, *si el proceso es aceptable,* es decir, que su producción no excede un 5% de artículos defectuosos, hay una probabilidad de 0.0282 de obtener 10 o más artículos defectuosos en una muestra aleatoria de 100 artículos del proceso. Esta pequeña probabilidad sugiere que, en realidad, a largo plazo el proceso tiene un porcentaje de artículos defectuosos mayor al 5%. En otras palabras, en las condiciones de un proceso aceptable casi nunca se obtendría la información muestral que se obtuvo. Sin embargo, ¡se obtuvo! Por lo tanto, es evidente que la probabilidad de que se obtuviera sería mucho mayor si la tasa de artículos defectuosos del proceso fuera mucho mayor que 5%. ■

A partir de este ejemplo se vuelve evidente que los elementos de probabilidad ayudan a traducir la información muestral en algo concluyente o no concluyente acerca del sistema científico. De hecho, lo aprendido probablemente constituya información inquietante para el ingeniero o administrador. Los métodos estadísticos (que examinaremos con más detalle en el capítulo 10) produjeron un valor-*P* de 0.0282. El resultado sugiere que es **muy probable que el proceso no sea aceptable**. En los capítulos siguientes se trata detenidamente el concepto de **valor-*P***. El próximo ejemplo brinda una segunda ilustración.

Ejemplo 1.2: Con frecuencia, la naturaleza del estudio científico señalará el papel que desempeñan la probabilidad y el razonamiento deductivo en la inferencia estadística. El ejercicio 9.40 en la página 294 proporciona datos asociados con un estudio que se llevó a cabo en el Virginia Polytechnic Institute and State University acerca del desarrollo de una relación entre las raíces de los árboles y la acción de un hongo. Los minerales de los hongos se transfieren a los árboles, y los azúcares de los árboles a los hongos. Se plantaron dos muestras de 10 plantones de roble rojo norteño en un invernadero, una de ellas contenía

plantones tratados con nitrógeno y la otra plantones sin tratamiento. Todas las demás condiciones ambientales se mantuvieron constantes. Todos los plantones contenían el hongo *Pisolithus tinctorus*. En el capítulo 9 se incluyen más detalles. Los pesos en gramos de los tallos se registraron después de 140 días y los datos se presentan en la tabla 1.1.

Tabla 1.1: Conjunto de datos del ejemplo 1.2

Sin nitrógeno	**Con nitrógeno**
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

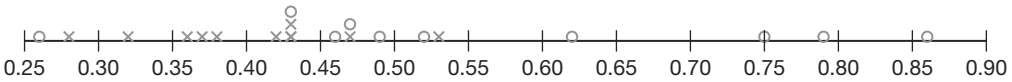


Figura 1.1: Gráfica de puntos de los datos de peso del tallo.

En este ejemplo hay dos muestras tomadas de dos **poblaciones distintas**. El objetivo del experimento es determinar si el uso del nitrógeno influye en el crecimiento de las raíces. Éste es un estudio comparativo (es decir, es un estudio en el que se busca comparar las dos poblaciones en cuanto a ciertas características importantes). Los datos se deben graficar como se indica en el diagrama de puntos de la figura 1.1. Los valores $\circ$ representan los datos "con nitrógeno" y los valores $\times$ los datos "sin nitrógeno".

Observe que la apariencia general de los datos podría sugerir al lector que, en promedio, el uso del nitrógeno aumenta el peso del tallo. Cuatro observaciones con nitrógeno son considerablemente más grandes que cualquiera de las observaciones sin nitrógeno. La mayoría de las observaciones sin nitrógeno parece estar por debajo del centro de los datos. La apariencia del conjunto de datos parece indicar que el nitrógeno es efectivo. Pero, ¿cómo se cuantifica esto? ¿Cómo se puede resumir toda la evidencia visual aparente de manera que tenga algún significado? Como en el ejemplo anterior, se pueden utilizar los fundamentos de la probabilidad. Las conclusiones se resumen en una declaración de probabilidad o valor-*P.* Aquí no demostraremos la inferencia estadística que produce la probabilidad resumida. Igual que en el ejemplo 1.1, tales métodos se estudiarán en el capítulo 10. El problema gira alrededor de la "probabilidad de que datos como éstos se puedan observar", *dado que el nitrógeno no tiene efecto*; en otras palabras, dado que ambas muestras se generaron a partir de la misma población. Suponga que esta probabilidad es pequeña, digamos de 0.03; un porcentaje que podría constituir suficiente evidencia de que el uso del nitrógeno en realidad influye en el peso promedio del tallo en los plantones de roble rojo (aparentemente lo aumenta). ∎

¿Cómo trabajan juntas la probabilidad y la inferencia estadística?

Es importante para el lector que comprenda claramente la diferencia entre la disciplina de la probabilidad, una ciencia por derecho propio, y la disciplina de la estadística inferencial. Como señalamos, el uso o la aplicación de conceptos de probabilidad permite interpretar la vida cotidiana a partir de los resultados de la inferencia estadística. En consecuencia, se afirma que la inferencia estadística emplea los conceptos de probabilidad. A partir de los dos ejemplos anteriores aprendimos que la información muestral está disponible para el analista y que, con la ayuda de métodos estadísticos y elementos de probabilidad, podemos obtener conclusiones acerca de alguna característica de la población (en el ejemplo 1.1 el proceso al parecer no es aceptable, y en el ejemplo 1.2 parece ser que el nitrógeno en verdad influye en el peso promedio de los tallos). Así, para un problema estadístico, **la muestra, junto con la estadística inferencial, nos permite obtener conclusiones acerca de la población, ya que la estadística inferencial utiliza ampliamente los elementos de probabilidad**. Tal razonamiento es *inductivo* por naturaleza. Ahora, cuando avancemos al capítulo 2 y los siguientes, el lector encontrará que, a diferencia de lo que hicimos en nuestros dos ejemplos actuales, no nos enfocaremos en resolver problemas estadísticos. En muchos de los ejemplos que estudiaremos no utilizaremos muestras. Lo que haremos será describir claramente una población con todas sus características conocidas. Las preguntas importantes se enfocarán en la naturaleza de los datos que hipotéticamente se podrían obtener a partir de la población. Entonces, podríamos afirmar que los **elementos de probabilidad nos permiten sacar conclusiones acerca de las características de los datos hipotéticos que se tomen de la población, con base en las características conocidas de la población**. Esta clase de razonamiento es *deductivo* por naturaleza. La figura 1.2 muestra la relación básica entre la probabilidad y la estadística inferencial.

Figura 1.2: Relación básica entre la probabilidad y la estadística inferencial.

Ahora bien, en términos generales, ¿cuál campo es más importante, el de la probabilidad o el de la estadística? Ambos son muy importantes y evidentemente se complementan. La única certeza respecto de la didáctica de ambas disciplinas radica en el hecho de que, si la estadística se debe enseñar con un nivel mayor al de un simple "libro de cocina", entonces hay que comenzar por enseñar la disciplina de la probabilidad. Esta regla se basa en el hecho de que un analista no podrá aprender nada sobre una población a partir de una muestra hasta que aprenda los rudimentos de incertidumbre en esa muestra. Considere el ejemplo 1.1; en el que la pregunta se centra en si la población, definida por el proceso, tiene o no más de 5% de elementos defectuosos. En otras palabras, la suposición es que 5 de cada 100 artículos, **en promedio,** salen defectuosos. Ahora bien, la muestra contiene 100 artículos y 10 están defectuosos. ¿Esto apoya o refuta la supo-

sición? Aparentemente la refuta porque 10 artículos de cada 100 parecen ser "un trozo grande". ¿Pero cómo podríamos saber esto sin tener nociones de probabilidad? La única manera en que podremos aprender las condiciones en las cuales el proceso es aceptable (5% de defectuosos) es estudiando el material de los siguientes capítulos. La probabilidad de obtener 10 o más artículos defectuosos en una muestra de 100 es de 0.0282.

Dimos dos ejemplos en donde los elementos de probabilidad ofrecen un resumen que el científico o el ingeniero pueden usar como evidencia para basar una decisión. El puente entre los datos y la conclusión está, por supuesto, basado en los fundamentos de la inferencia estadística, la teoría de la distribución y las distribuciones de muestreos que se examinarán en capítulos posteriores.

1.2 Procedimientos de muestreo; recolección de los datos

En la sección 1.1 estudiamos muy brevemente el concepto de muestreo y el proceso de muestreo. Aunque el muestreo parece ser un concepto simple, la complejidad de las preguntas que se deben contestar acerca de la población, o las poblaciones, en ocasiones requiere que el proceso de muestreo sea muy complejo. El concepto de muestreo se examinará de manera técnica en el capítulo 8, pero aquí nos esforzaremos por dar algunas nociones de sentido común sobre el muestreo. Ésta es una transición natural hacia el análisis del concepto de variabilidad.

Muestreo aleatorio simple

La importancia del muestreo adecuado gira en torno al grado de confianza con que el analista es capaz de responder las preguntas que se plantean. Supongamos que sólo hay una población en el problema. Recuerde que en el ejemplo 1.2 había dos poblaciones implicadas. El **muestreo aleatorio simple** significa que cierta muestra dada de un *tamaño muestral* específico tiene la misma probabilidad de ser seleccionada que cualquiera otra muestra del mismo tamaño. El término **tamaño muestral** simplemente indica el número de elementos en la muestra. Evidentemente, en muchos casos se puede utilizar una tabla de números aleatorios para seleccionar la muestra. La ventaja del muestreo aleatorio simple radica en que ayuda a eliminar el problema de tener una muestra que refleje una población diferente (quizá más restringida) de aquella sobre la cual se necesitan realizar las inferencias. Por ejemplo, se elige una muestra para contestar diferentes preguntas respecto de las preferencias políticas en cierta entidad de Estados Unidos. La muestra implica la elección de, digamos, 1 000 familias y una encuesta a aplicar. Ahora bien, suponga que no se utiliza el muestreo aleatorio, sino que todas o casi todas las 1 000 familias se eligen de una zona urbana. Se considera que las preferencias políticas en las áreas rurales difieren de las de las áreas urbanas. En otras palabras, la muestra obtenida en realidad confinó a la población y, por lo tanto, las inferencias también se tendrán que restringir a la "población confinada", y en este caso el confinamiento podría resultar indeseable. Si, de hecho, se necesitara hacer las inferencias respecto de la entidad como un todo, a menudo se diría que la muestra con un tamaño de 1 000 familias aquí descrita es una **muestra sesgada**.

Como antes sugerimos, el muestreo aleatorio simple no siempre es adecuado. El enfoque alternativo que se utilice dependerá de la complejidad del problema. Con frecuencia, por ejemplo, las unidades muestrales no son homogéneas y se dividen naturalmente en grupos que no se traslapan y que son homogéneos. Tales grupos se llaman *estratos*, y

un procedimiento llamado *muestreo aleatorio estratificado* implica la selección al azar de una muestra *dentro* de cada estrato. El propósito de esto es asegurarse de que ninguno de los estratos esté sobrerrepresentado ni subrepresentado. Por ejemplo, suponga que se aplica una encuesta a una muestra para reunir opiniones preliminares respecto de un referéndum que se piensa realizar en determinada ciudad. La ciudad está subdividida en varios grupos étnicos que representan estratos naturales y, para no excluir ni sobrerrepresentar a algún grupo de cada uno de ellos, se eligen muestras aleatorias separadas de cada grupo.

Diseño experimental

El concepto de aleatoriedad o asignación aleatoria desempeña un papel muy importante en el área del **diseño experimental**, que se presentó brevemente en la sección 1.1 y es un fundamento muy importante en casi cualquier área de la ingeniería y de la ciencia experimental. Estudiaremos este tema con detenimiento en los capítulos 13 a 15. Sin embargo, es conveniente introducirlo aquí brevemente en el contexto del muestreo aleatorio. Un conjunto de los llamados **tratamientos** o **combinaciones de tratamientos** se vuelven las poblaciones que se van a estudiar o a comparar en algún sentido. Un ejemplo es el tratamiento "con nitrógeno" *versus* "sin nitrógeno" del ejemplo 1.2. Otro ejemplo sencillo sería "placebo" *versus* "medicamento activo" o, en un estudio sobre la fatiga por corrosión, tendríamos combinaciones de tratamientos que impliquen especímenes con recubrimiento o sin recubrimiento, así como condiciones de alta o de baja humedad, a las cuales se somete el espécimen. De hecho, habrían cuatro combinaciones de factores o de tratamientos (es decir, 4 poblaciones), y se podrían formular y responder muchas preguntas científicas usando los métodos estadísticos e inferenciales. Considere primero la situación del ejemplo 1.2. En el experimento hay 20 plantones enfermos implicados. A partir de los datos es fácil observar que los plantones son diferentes entre sí. Dentro del grupo tratado con nitrógeno (o del grupo que no se trató con nitrógeno) hay **variabilidad** considerable en el peso de los tallos, la cual se debe a lo que por lo general se denomina **unidad experimental**. Éste es un concepto tan importante en la estadística inferencial que no es posible describirlo totalmente en este capítulo. La naturaleza de la variabilidad es muy importante. Si es demasiado grande, debido a que resulta de una condición de excesiva falta de homogeneidad en las unidades experimentales, la variabilidad "eliminará" cualquier diferencia detectable entre ambas poblaciones. Recuerde que en este caso eso no ocurrió.

La gráfica de puntos de la figura 1.1 y el valor-*P* indican una clara distinción entre esas dos condiciones. ¿Qué papel desempeñan tales unidades experimentales en el proceso mismo de recolección de los datos? El enfoque por sentido común y, de hecho, estándar, es asignar los 20 plantones o unidades experimentales **aleatoriamente a las dos condiciones o tratamientos**. En el estudio del medicamento podríamos decidir utilizar un total de 200 pacientes disponibles, quienes serán claramente distinguibles en algún sentido. Ellos son las unidades experimentales. No obstante, tal vez todos tengan una condición crónica que podría ser tratada con el fármaco. Así, en el denominado **diseño completamente aleatorio**, se asignan al azar 100 pacientes al placebo y 100 al medicamento activo. De nuevo, son estas unidades experimentales en el grupo o tratamiento las que producen la variabilidad en el resultado de los datos (es decir, la variabilidad en el resultado medido), digamos, de la presión sanguínea o cualquier valor de la eficacia de un medicamento que sea importante. En el estudio de la fatiga por corrosión las unidades experimentales son los especímenes que se someten a la corrosión.

¿Por qué las unidades experimentales se asignan aleatoriamente?

¿Cuál es el posible efecto negativo de no asignar aleatoriamente las unidades experimentales a los tratamientos o a las combinaciones de tratamientos? Esto se observa más claramente en el caso del estudio del medicamento. Entre las características de los pacientes que producen variabilidad en los resultados están la edad, el género y el peso. Tan sólo suponga que por casualidad el grupo del placebo contiene una muestra de personas que son predominantemente más obesas que las del grupo del tratamiento. Quizá los individuos más obesos muestren una tendencia a tener una presión sanguínea más elevada, lo cual evidentemente sesgará el resultado y, por lo tanto, cualquier resultado que se obtenga al aplicar la inferencia estadística podría tener poco que ver con el efecto del medicamento, pero mucho con las diferencias en el peso de ambas muestras de pacientes.

Deberíamos enfatizar la importancia del término **variabilidad**. La variabilidad excesiva entre las unidades experimentales "disfraza" los hallazgos científicos. En secciones posteriores intentaremos clasificar y cuantificar las medidas de variabilidad. En las siguientes secciones presentaremos y analizaremos cantidades específicas que se calculan en las muestras; las cantidades proporcionan una idea de la naturaleza de la muestra respecto de la ubicación del centro de los datos y la variabilidad de los mismos. Un análisis de varias de tales medidas de un solo número permite ofrecer un preámbulo de que la información estadística será un componente importante de los métodos estadísticos que se utilizarán en capítulos posteriores. Estas medidas, que ayudan a clasificar la naturaleza del conjunto de datos, caen en la categoría de **estadísticas descriptivas**. Este material es una introducción a una presentación breve de los métodos pictóricos y gráficos que van incluso más allá en la caracterización del conjunto de datos. El lector debería entender que los métodos estadísticos que se presentan aquí se utilizarán a lo largo de todo el texto. Para ofrecer una imagen más clara de lo que implican los estudios de diseño experimental se presenta el ejemplo 1.3.

Ejemplo 1.3: Se realizó un estudio sobre la corrosión con la finalidad de determinar si al recubrir una aleación de aluminio con una sustancia retardadora de la corrosión, el metal se corroe menos. El recubrimiento es un protector que los anunciantes afirman que minimiza el daño por fatiga en esta clase de material. La influencia de la humedad sobre la magnitud de la corrosión también es de interés. Una medición de la corrosión puede expresarse en millares de ciclos hasta la ruptura del metal. Se utilizaron dos niveles de recubrimiento: sin recubrimiento y con recubrimiento químico contra la corrosión. También se consideraron dos niveles de humedad relativa, de 20% y 80%, respectivamente.

El experimento implica las cuatro combinaciones de tratamientos que se listan en la siguiente tabla. Se usan ocho unidades experimentales, que son especímenes de aluminio preparados, dos de los cuales se asignan aleatoriamente a cada una de las cuatro combinaciones de tratamiento. Los datos se presentan en la tabla 1.2.

Los datos de la corrosión son promedios de los dos especímenes. En la figura 1.3 se presenta una gráfica con los promedios. Un valor relativamente grande de ciclos hasta la ruptura representa una cantidad pequeña de corrosión. Como se podría esperar, al parecer un incremento en la humedad hace que empeore la corrosión. El uso del procedimiento de recubrimiento químico contra la corrosión parece reducir la corrosión. ∎

En este ejemplo de diseño experimental el ingeniero eligió sistemáticamente las cuatro combinaciones de tratamiento. Para vincular esta situación con los conceptos con los que el lector se ha familiarizado hasta aquí, deberíamos suponer que las condiciones

Tabla 1.2: Datos para el ejemplo 1.3

Recubrimiento	Humedad	Promedio de corrosión en miles de ciclos hasta la ruptura
Sin recubrimiento	20%	975
	80%	350
Con recubrimiento químico contra la corrosión	20%	1750
	80%	1550

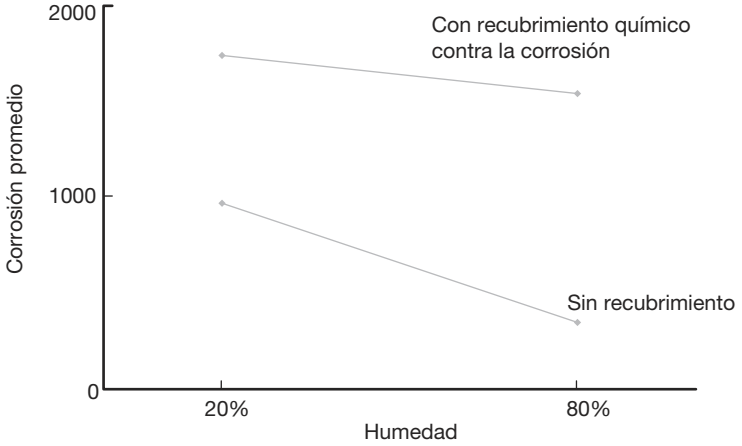


Figura 1.3: Resultados de corrosión para el ejemplo 1.3.

que representan las cuatro combinaciones de tratamientos son cuatro poblaciones separadas y que los dos valores de corrosión observados en cada una de las poblaciones constituyen importantes piezas de información. La importancia del promedio al captar y resumir ciertas características en la población se destacará en la sección 1.3. Aunque a partir de la figura podríamos sacar conclusiones acerca del papel que desempeña la humedad y del efecto de recubrir los especímenes, no podemos evaluar con exactitud los resultados de un punto de vista analítico sin tomar en cuenta la *variabilidad alrededor* del promedio. De nuevo, como señalamos con anterioridad, si los dos valores de corrosión en cada una de las combinaciones de tratamientos son muy cercanos, la imagen de la figura 1.3 podría ser una descripción precisa. Pero si cada valor de la corrosión en la figura es un promedio de dos valores que están ampliamente dispersos, entonces esta variabilidad podría, de hecho, en verdad "eliminar" cualquier información que parezca difundirse cuando tan sólo se observan los promedios. Los siguientes ejemplos ilustran estos conceptos:

1. La asignación aleatoria a las combinaciones de tratamientos (recubrimiento/humedad) de las unidades experimentales (especímenes).
2. El uso de promedios muestrales (valores de corrosión promedio) para resumir la información muestral.
3. La necesidad de considerar las medidas de variabilidad en el análisis de cualquier muestra o conjunto de muestras.

Este ejemplo sugiere la necesidad de estudiar el tema que se expone en las secciones 1.3 y 1.4, es decir, el de las estadísticas descriptivas que indican las medidas de la ubicación del centro en un conjunto de datos, y aquellas con las que se mide la variabilidad.

1.3 Medidas de localización: la media y la mediana de una muestra

Las medidas de localización están diseñadas para brindar al analista algunos valores cuantitativos de la ubicación central o de otro tipo de los datos en una muestra. En el ejemplo 1.2 parece que el centro de la muestra con nitrógeno claramente excede al de la muestra sin nitrógeno. Una medida obvia y muy útil es la **media de la muestra**. La media es simplemente un promedio numérico.

Definición 1.1: Suponga que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$. La **media de la muestra**, que se denota con $\bar{x}$, es

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Hay otras medidas de tendencia central que se explican con detalle en capítulos posteriores. Una medida importante es la **mediana de la muestra**. El propósito de la mediana de la muestra es reflejar la tendencia central de la muestra de manera que no sea influida por los valores extremos.

Definición 1.2: Dado que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$, acomodadas en **orden de magnitud creciente**, la mediana de la muestra es

$$\tilde{x} = \begin{cases} x_{(n+1)/2,} & \text{si } n \text{ es impar,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$

Por ejemplo, suponga que el conjunto de datos es el siguiente: 1.7, 2.2, 3.9, 3.11 y 14.7. La media y la mediana de la muestra son, respectivamente,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Es evidente que la media es influida de manera considerable por la presencia de la observación extrema, 14.7; en tanto que el lugar de la mediana hace énfasis en el verdadero "centro" del conjunto de datos. En el caso del conjunto de datos de dos muestras del ejemplo 1.2, las dos medidas de tendencia central para las muestras individuales son

$$\begin{aligned} \bar{x} \text{ (sin nitrógeno)} &= 0.399 \text{ gramos,} \\ \tilde{x} \text{ (sin nitrógeno)} &= \frac{0.38 + 0.42}{2} = 0.400 \text{ gramos,} \\ \bar{x} \text{ (con nitrógeno)} &= 0.565 \text{ gramos,} \\ \tilde{x} \text{ (con nitrógeno)} &= \frac{0.49 + 0.52}{2} = 0.505 \text{ gramos.} \end{aligned}$$

Es evidente que hay una diferencia conceptual entre la media y la mediana. Para el lector con ciertas nociones de ingeniería quizá sea de interés que la media de la muestra

es el **centroide de los datos** en una muestra. En cierto sentido es el punto en el cual se puede colocar un fulcro (apoyo) para equilibrar un sistema de "pesos", que son las ubicaciones de los datos individuales. Esto se muestra en la figura 1.4 respecto de la muestra "con nitrógeno".

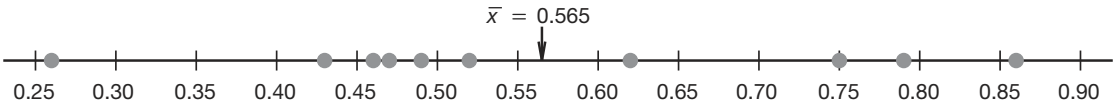


Figura 1.4: Media de la muestra como centroide del peso del tallo con nitrógeno.

En capítulos posteriores la base para el cálculo de $\bar{x}$ es un **estimado** de la **media de la población**. Como antes señalamos, el propósito de la inferencia estadística es obtener conclusiones acerca de las características o **parámetros** y la **estimación** es una característica muy importante de la inferencia estadística.

La mediana y la media pueden ser muy diferentes entre sí. Observe, sin embargo, que en el caso de los datos del peso de los tallos el valor de la media de la muestra para "sin nitrógeno" es bastante similar al valor de la mediana.

Otras medidas de localización

Hay muchos otros métodos para calcular la ubicación del centro de los datos en la muestra. No los trataremos en este momento. Por lo general las alternativas para la media de la muestra se diseñan con el fin de generar valores que representen relación entre la media y la mediana. Rara vez utilizamos alguna de tales medidas. Sin embargo, es aleccionador estudiar una clase de estimadores conocida como **media recortada**, la cual se calcula "quitando" cierto porcentaje de los valores mayores y menores del conjunto. Por ejemplo, la media recortada al 10% se encuentra eliminando tanto el 10% de los valores mayores como el 10% de los menores, y calculando el promedio de los valores restantes. En el caso de los datos del peso de los tallos, eliminaríamos el valor más alto y el más bajo, ya que el tamaño de la muestra es 10 en cada caso. De manera que para el grupo sin nitrógeno la media recortada al 10% está dada por

$$\bar{x}_{\text{rec (10)}} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

y para la media recortada al 10% del grupo con nitrógeno tenemos

$$\bar{x}_{\text{rec (10)}} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Observe que en este caso, como se esperaba, las medias recortadas están cerca tanto de la media como de la mediana para las muestras individuales. Desde luego, el enfoque de la media recortada es menos sensible a los valores extremos que la media de la muestra, pero no tan insensible como la mediana. Además, el método de la media recortada utiliza más información que la mediana de la muestra. Advierta que la mediana de la muestra es, de hecho, un caso especial de la media recortada, en el cual se eliminan todos los datos de la muestra y queda sólo el central o dos observaciones.

Ejercicios

1.1 Se registran las siguientes mediciones para el tiempo de secado (en horas) de cierta marca de pintura esmaltada.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones constituyen una muestra aleatoria simple.

a) ¿Cuál es el tamaño de la muestra anterior?
b) Calcule la media de la muestra para estos datos.
c) Calcule la mediana de la muestra.
d) Grafique los datos utilizando una gráfica de puntos.
e) Calcule la media recortada al 20% para el conjunto de datos anterior.
f) ¿La media muestral para estos datos es más o menos descriptiva como centro de localización, que la media recortada?

1.2 Según la revista *Chemical Engineering*, una propiedad importante de una fibra es su absorción del agua. Se toma una muestra aleatoria de 20 pedazos de fibra de algodón y se mide la absorción de cada uno. Los valores de absorción son los siguientes:

18.71	21.41	20.72	21.81	19.29	22.43	20.17
23.71	19.44	20.50	18.92	20.33	23.00	22.85
19.25	21.77	22.11	19.77	18.04	21.12	

a) Calcule la media y la mediana muestrales para los valores de la muestra anterior.
b) Calcule la media recortada al 10%.
c) Elabore una gráfica de puntos con los datos de la absorción.
d) Si se utilizan sólo los valores de la media, la mediana y la media recortada, ¿hay evidencia de valores extremos en los datos?

1.3 Se utiliza cierto polímero para los sistemas de evacuación de los aviones. Es importante que el polímero sea resistente al proceso de envejecimiento. Se utilizaron veinte especímenes del polímero en un experimento. Diez se asignaron aleatoriamente para exponerse a un proceso de envejecimiento acelerado del lote, el cual implica la exposición a altas temperaturas durante 10 días. Se hicieron las mediciones de resistencia a la tensión de los especímenes y se registraron los siguientes datos sobre resistencia a la tensión en psi.

Sin envejecimiento:	227	222	218	217	225
	218	216	229	228	221
Con envejecimiento:	219	214	215	211	209
	218	203	204	201	205

a) Elabore la gráfica de puntos de los datos.
b) ¿En la gráfica que obtuvo parece que el proceso de envejecimiento tuvo un efecto en la resistencia a la tensión de este polímero? Explique su respuesta.
c) Calcule la resistencia a la tensión de la media de la muestra en las dos muestras.
d) Calcule la mediana de ambas. Analice la similitud o falta de similitud entre la media y la mediana de cada grupo.

1.4 En un estudio realizado por el Departamento de Ingeniería Mecánica del Tecnológico de Virginia se compararon las varillas de acero que abastecen dos compañías diferentes. Se fabricaron diez resortes de muestra con las varillas de metal proporcionadas por cada una de las compañías y se registraron sus medidas de flexibilidad. Los datos son los siguientes:

Compañía A:	9.3	8.8	6.8	8.7	8.5
	6.7	8.0	6.5	9.2	7.0
Compañía B:	11.0	9.8	9.9	10.2	10.1
	9.7	11.0	11.1	10.2	9.6

a) Calcule la media y la mediana de la muestra para los datos de ambas compañías.
b) Grafique los datos para las dos compañías en la misma línea y explique su conclusión respecto de cualquier aparente diferencia entre las dos compañías.

1.5 Veinte hombres adultos de entre 30 y 40 años de edad participaron en un estudio para evaluar el efecto de cierto régimen de salud, que incluye dieta y ejercicio, en el colesterol sanguíneo. Se eligieron aleatoriamente diez para el grupo de control y los otros diez se asignaron para participar en el régimen como el grupo de tratamiento durante un periodo de seis meses. Los siguientes datos muestran la reducción en el colesterol que experimentaron en ese periodo los 20 sujetos:

Grupo de control:	7	3	−4	14	2
	5	22	−7	9	5
Grupo de tratamiento:	−6	5	9	4	4
	12	37	5	3	3

a) Elabore una gráfica de puntos con los datos de ambos grupos en la misma gráfica.
b) Calcule la media, la mediana y la media recortada al 10% para ambos grupos.
c) Explique por qué la diferencia en las medias sugiere una conclusión acerca del efecto del régimen, en tanto que la diferencia en las medianas o las medias recortadas sugiere una conclusión diferente.

1.6 La resistencia a la tensión del caucho de silicio se considera una función de la temperatura de vulcanizado. Se llevó a cabo un estudio en el que se prepararon muestras de 12 especímenes del caucho utilizando temperaturas de vulcanizado de 20°C y 45°C. Los siguientes

datos presentan los valores de resistencia a la tensión en megapascales.

20°C:	2.07	2.14	2.22	2.03	2.21	2.03
	2.05	2.18	2.09	2.14	2.11	2.02
45°C:	2.52	2.15	2.49	2.03	2.37	2.05
	1.99	2.42	2.08	2.42	2.29	2.01

a) Elabore una gráfica de puntos con los datos, tanto de los valores de resistencia a la tensión a temperatura alta como los de a temperatura baja.

b) Calcule la resistencia a la tensión media muestral para ambas muestras.

c) Al observar la gráfica, ¿le parece que la temperatura de vulcanizado influye en la resistencia a la tensión? Explique su respuesta.

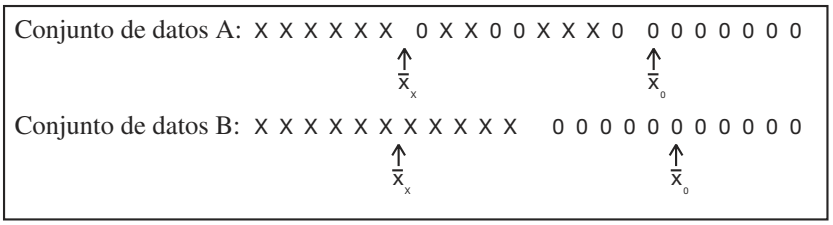
d) ¿En qué otra cosa, al parecer, influye el incremento en la temperatura de vulcanizado? Explique su respuesta.

1.4 Medidas de variabilidad

La variabilidad de una muestra desempeña un papel importante en el análisis de datos. La variabilidad de procesos y productos es un hecho real en los sistemas científicos y de ingeniería: el control o la reducción de la variabilidad de un proceso a menudo es una fuente de mayores dificultades. Cada vez más ingenieros y administradores de procesos están aprendiendo que la calidad del producto y, como resultado, las utilidades que se derivan de los productos manufacturados es, con mucho, una función de la **variabilidad del proceso**. En consecuencia, gran parte de los capítulos 9 a 15 se dedica al análisis de datos y a los procedimientos de modelado en los que la variabilidad de la muestra desempeña un papel significativo. Incluso en problemas pequeños de análisis de datos el éxito de un método estadístico específico podría depender de la magnitud de la variabilidad entre las observaciones en la muestra. Las medidas de ubicación en una muestra no brindan un resumen adecuado de la naturaleza de un conjunto de datos. Considere el ejemplo 1.2, en el que no podemos concluir que el uso del nitrógeno aumenta el crecimiento sin tomar en cuenta la variabilidad de la muestra.

Aunque los detalles del análisis de este tipo de conjuntos de datos se estudiarán en el capítulo 9, a partir de la figura 1.1 debería quedar claro que la variabilidad entre las observaciones sin nitrógeno y la variabilidad entre las observaciones con nitrógeno tiene, desde luego, alguna consecuencia. De hecho, parece que la variabilidad dentro de la muestra con nitrógeno es mayor que la de la muestra sin nitrógeno. Quizás haya algo acerca de la inclusión del nitrógeno que no sólo incrementa el peso de los tallos ($\bar{x}$ de 0.565 gramos en comparación con una $\bar{x}$ de 0.399 gramos para la muestra sin nitrógeno), sino que también incrementa la variabilidad en el peso de los tallos (es decir, provoca que el peso de los tallos sea más inconsistente).

Por ejemplo, compare los dos conjuntos de datos de abajo. Cada uno contiene dos muestras y la diferencia en las medias es aproximadamente la misma para ambas, aunque el conjunto de datos B parece proporcionar un contraste mucho más claro entre las dos poblaciones de las que se tomaron las muestras. Si el propósito de tal experimento es detectar la diferencia entre las dos poblaciones, esto se logra en el caso del conjunto de datos B. Sin embargo, en el conjunto de datos A la amplia variabilidad *dentro* de las dos muestras ocasiona dificultad. De hecho, no es claro que haya una diferencia *entre* las dos poblaciones.



Rango y desviación estándar de la muestra

Así como hay muchas medidas de tendencia central o de localización, hay muchas medidas de dispersión o variabilidad. Quizá la más simple sea el **rango de la muestra** $X_{\text{máx}} - X_{\text{mín}}$. El rango puede ser muy útil y se examina con amplitud en el capítulo 17 sobre *control estadístico de calidad.* La medida muestral de dispersión que se utiliza más a menudo es la **desviación estándar de la muestra**. Nuevamente denotemos con x_1, $x_2, \ldots, x_n$ los valores de la muestra.

Definición 1.3: La **varianza de la muestra**, denotada con s^2, está dada por

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}.$$

La **desviación estándar de la muestra**, denotada con s, es la raíz cuadrada positiva de s^2, es decir,

$$s = \sqrt{s^2}.$$

Para el lector debería quedar claro que la desviación estándar de la muestra es, de hecho, una medida de variabilidad. Una variabilidad grande en un conjunto de datos produce valores relativamente grandes de $(x - \bar{x})^2$ y, por consiguiente, una varianza muestral grande. La cantidad $n - 1$ a menudo se denomina **grados de libertad asociados con la varianza** estimada. En este ejemplo sencillo los grados de libertad representan el número de piezas de información independientes disponibles para calcular la variabilidad. Por ejemplo, suponga que deseamos calcular la varianza de la muestra y la desviación estándar del conjunto de datos (5, 17, 6, 4). El promedio de la muestra es $\bar{x} = 8$. El cálculo de la varianza implica:

$$(5 - 8)^2 + (17 - 8)^2 + (6 - 8)^2 + (4 - 8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

Las cantidades dentro de los paréntesis suman cero. En general, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (véase el ejercicio 1.16 de la página 31). Entonces, el cálculo de la varianza de una muestra no implica n **desviaciones cuadradas independientes** de la media $\bar{x}$. De hecho, como el último valor de $x - \bar{x}$ es determinado por los primeros $n - 1$ valores, decimos que éstas son $n - 1$ "piezas de información" que produce s^2. Por consiguiente, hay $n - 1$ grados de libertad en vez de n grados de libertad para calcular la varianza de una muestra.

Ejemplo 1.4: En un ejemplo que se estudia ampliamente en el capítulo 10, un ingeniero se interesa en probar el "sesgo" en un medidor de pH. Los datos se recaban con el medidor mediante la medición del pH de una sustancia neutra (pH = 7.0). Se toma una muestra de tamaño 10 y se obtienen los siguientes resultados:

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.

La media de la muestra $\bar{x}$ está dada por

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \ldots + 7.08}{10} = 7.0250.$$

La varianza de la muestra s^2 está dada por

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

Como resultado, la desviación estándar de la muestra está dada por

$$s = \sqrt{0.001939} = 0.044.$$

Así que la desviación estándar de la muestra es 0.0440 con $n - 1 = 9$ grados de libertad. ∎

Unidades para la desviación estándar y la varianza

A partir de la definición 1.3 debería ser evidente que la varianza es una medida de la desviación cuadrática promedio de la media $\bar{x}$. Empleamos el término *desviación cuadrática promedio* aun cuando la definición utilice una división entre $n - 1$ grados de libertad en vez de n. Desde luego, si n es grande, la diferencia en el denominador es inconsecuente. Por lo tanto, la varianza de la muestra tiene unidades que son el cuadrado de las unidades en los datos observados; aunque la desviación estándar de la muestra se encuentra en unidades lineales. Considere los datos del ejemplo 1.2. Los pesos del tallo se miden en gramos. Como resultado, las desviaciones estándar de la muestra están en gramos y las varianzas se miden en gramos^2. De hecho, las desviaciones estándar individuales son 0.0728 gramos para el caso sin nitrógeno y 0.1867 gramos para el grupo con nitrógeno. Observe que la desviación estándar en verdad indica una variabilidad mucho más grande en la muestra con nitrógeno. Esta condición se destaca en la figura 1.1.

¿Cuál es la medida de variabilidad más importante?

Como indicamos antes, el rango de la muestra tiene aplicaciones en el área del control estadístico de la calidad. Quizás el lector considere que es redundante utilizar la varianza de la muestra y la desviación estándar de la muestra. Ambas medidas reflejan el mismo concepto en la variabilidad de la medición, pero la desviación estándar de la muestra mide la variabilidad en unidades lineales; en tanto que la varianza muestral se mide en unidades cuadradas. Ambas desempeñan papeles importantes en el uso de los métodos estadísticos. Mucho de lo que se logra en el contexto de la inferencia estadística implica la obtención de conclusiones acerca de las características de poblaciones. Entre tales características son constantes los denominados **parámetros de la población**. Dos parámetros importantes son la **media de la población** y la **varianza de la población**. La varianza de la muestra desempeña un papel explícito en los métodos estadísticos que se utilizan para obtener inferencias sobre la varianza de la población. La desviación estándar de la muestra desempeña un papel importante, junto con la media de la muestra, en las inferencias que se realizan acerca de la media de la población. En general, la varianza se considera más en la teoría inferencial, mientras que la desviación estándar se utiliza más en aplicaciones.

Ejercicios

1.7 Considere los datos del tiempo de secado del ejercicio 1.1 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra.

1.8 Calcule la varianza de la muestra y la desviación estándar para los datos de absorción del agua del ejercicio 1.2 de la página 13.

1.9 El ejercicio 1.3 de la página 13 presentó datos de resistencia a la tensión de dos muestras, una en la que los especímenes se expusieron a un proceso de envejecimiento y otra en la que no se efectuó tal proceso en los especímenes.

a) Calcule la varianza de la muestra, así como su desviación estándar, en cuanto a la resistencia a la tensión en ambas muestras.

b) ¿Parece haber alguna evidencia de que el envejecimiento afecta la variabilidad en la resistencia a la tensión? (Véase también la gráfica para el ejercicio 1.3 de la página 13).

1.10 Para los datos del ejercicio 1.4 de la página 13 calcule tanto la media como la varianza de la "flexibilidad" para las compañías A y B. ¿Parece que hay una diferencia de flexibilidad entre la compañía A y la compañía B?

1.11 Considere los datos del ejercicio 1.5 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra para ambos grupos: el de tratamiento y el de control.

1.12 Para el ejercicio 1.6 de la página 13 calcule la desviación estándar muestral de la resistencia a la tensión para las muestras, de forma separada para ambas temperaturas. ¿Parece que un incremento en la temperatura influye en la variabilidad de la resistencia a la tensión? Explique su respuesta.

1.5 Datos discretos y continuos

La inferencia estadística a través del análisis de estudios observacionales o de diseños experimentales se utiliza en muchas áreas científicas. Los datos reunidos pueden ser **discretos** o **continuos**, según el área de aplicación. Por ejemplo, un ingeniero químico podría estar interesado en un experimento que lo lleve a condiciones en que se maximice la producción. Aquí, por supuesto, la producción se expresaría en porcentaje, o gramos/libra, medida en un continuo. Por otro lado, un toxicólogo que realice un experimento de combinación de fármacos quizás encuentre datos que son binarios por naturaleza (es decir, el paciente responde o no lo hace).

En la teoría de la probabilidad se hacen distinciones importantes entre datos discretos y continuos que nos permiten hacer inferencias estadísticas. Con frecuencia las aplicaciones de la inferencia estadística se encuentran cuando se trabaja con *datos por conteo*. Por ejemplo, un ingeniero podría estar interesado en estudiar el número de partículas radiactivas que pasan a través de un contador en, digamos, 1 milisegundo. Al personal responsable de la eficiencia de una instalación portuaria podría interesarle conocer las características del número de buques petroleros que llegan diariamente a cierta ciudad portuaria. En el capítulo 5 se examinarán varios escenarios diferentes que conducen a distintas formas de manejo de los datos para situaciones con datos por conteo.

Incluso en esta fase inicial del texto se debería poner especial atención a algunos detalles que se asocian con datos binarios. Son muchas las aplicaciones que requieren el análisis estadístico de datos binarios. Con frecuencia la medición que se utiliza en el análisis es la *proporción muestral*. En efecto, la situación binaria implica dos categorías. Si en los datos hay n unidades y x se define como el número que cae en la categoría 1, entonces $n - x$ cae en la categoría 2. Así, x/n es la proporción muestral en la categoría 1 y $1 - x/n$ es la proporción muestral en la categoría 2. En la aplicación biomédica, por ejemplo, 50 pacientes representarían las unidades de la muestra y si, después de que se les suministra el medicamento, 20 de los 50 experimentaran mejoría en sus malestares estomacales (que son comunes en los 50), entonces $\frac{20}{50} = 0.4$ sería la proporción muestral

para la cual el medicamento tuvo éxito, y $1 - 0.4 = 0.6$ sería la proporción muestral para la cual el fármaco no tuvo éxito. En realidad, la medición numérica fundamental para datos binarios por lo general se denota con 0 o con 1. Éste es el caso de nuestro ejemplo médico, en el que un resultado exitoso se denota con un 1 y uno no exitoso con un 0. Entonces, la proporción muestral es en realidad una media muestral de unos y ceros. Para la categoría de éxitos,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

¿Qué clases de problemas se resuelven en situaciones con datos binarios?

Los tipos de problemas que enfrentan científicos e ingenieros que usan datos binarios no son muy difíciles, a diferencia de aquellos en los que las mediciones de interés son las continuas. Sin embargo, se utilizan técnicas diferentes debido a que las propiedades estadísticas de las proporciones muestrales son bastante diferentes de las medias muestrales que resultan de los promedios tomados de poblaciones continuas. Considere los datos del ejemplo en el ejercicio 1.6 de la página 13. El problema estadístico subyacente en este caso se enfoca en si una intervención, digamos un incremento en la temperatura de vulcanizado, alterará la resistencia a la tensión de la media de la población que se asocia con el proceso del caucho de silicio. Por otro lado, en el área de control de calidad, suponga que un fabricante de neumáticos para automóvil informa que en un embarque con 5000 neumáticos, seleccionados aleatoriamente del proceso, hay 100 defectuosos. Aquí la proporción muestral es $\frac{100}{5000} = 0.02$. Luego de realizar un cambio en el proceso diseñado para reducir los neumáticos defectuosos, se toma una segunda muestra de 5000 y se encuentran 90 defectuosos. La proporción muestral se redujo a $\frac{90}{5000} = 0.018$. Entonces, surge una pregunta: "¿La disminución en la proporción muestral de 0.02 a 0.018 es suficiente para sugerir una mejoría real en la proporción de la población?" En ambos casos se requiere el uso de las propiedades estadísticas de los promedios de la muestra: uno de las muestras de poblaciones continuas y el otro de las muestras de poblaciones discretas (binarias). En ambos casos la media de la muestra es un **estimado** de un parámetro de la población, una media de la población en el primer caso (es decir, la media de la resistencia a la tensión) y una proporción de la población (o sea, la proporción de neumáticos defectuosos en la población) en el segundo caso. Así que aquí tenemos estimados de la muestra que se utilizan para obtener conclusiones científicas respecto de los parámetros de la población. Como indicamos en la sección 1.3, éste es el tema general en muchos problemas prácticos en los que se usa la inferencia estadística.

1.6 Modelado estadístico, inspección científica y diagnósticos gráficos

A menudo, el resultado final de un análisis estadístico es la estimación de los parámetros de un **modelo postulado.** Éste es un proceso natural para los científicos y los ingenieros, ya que con frecuencia usan modelos. Un modelo estadístico no es determinista, es más bien un modelo que conlleva algunos aspectos probabilísticos. A menudo una forma de modelo es la base de las **suposiciones** que hace el analista. En el ejemplo 1.2 el científico podría desear determinar, a través de la información de la muestra, algún nivel de distinción entre las poblaciones tratadas con nitrógeno y las poblaciones no tratadas. El análisis podría requerir cierto modelo para los datos; por ejemplo, que las dos muestras

provengan de **distribuciones normales** o **gaussianas**. Véase el capítulo 6 para el estudio de la distribución normal.

Es evidente que quienes utilizan métodos estadísticos no pueden generar la información o los datos experimentales suficientes para describir a la totalidad de la población. Pero es frecuente que se utilicen los conjuntos de datos para aprender sobre ciertas propiedades de la población. Los científicos y los ingenieros están acostumbrados a manejar conjuntos de datos. Debería ser obvia la importancia de describir o *resumir* la naturaleza de los conjuntos de datos. Con frecuencia el resumen gráfico de un conjunto de datos puede proporcionar información sobre el sistema del que se obtuvieron los datos. Por ejemplo, en las secciones 1.1 y 1.3 mostramos gráficas de puntos.

En esta sección se estudia con detalle el papel del muestreo y de la graficación de los datos para mejorar la **inferencia estadística**. Nos limitamos a presentar algunas gráficas sencillas, pero a menudo efectivas, que complementan el estudio de poblaciones estadísticas.

Diagrama de dispersión

A veces el modelo postulado puede tener una forma algo más compleja. Por ejemplo, considere a un fabricante de textiles que diseña un experimento en donde se producen especímenes de tela que contienen diferentes porcentajes de algodón. Considere los datos de la tabla 1.3.

Tabla 1.3: Resistencia a la tensión

Porcentaje del algodón	Resistencia a la tensión
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Se fabrican cinco especímenes de tela para cada uno de los cuatro porcentajes de algodón. En este caso tanto el modelo para el experimento como el tipo de análisis que se utiliza deberían tomar en cuenta el objetivo del experimento y los insumos importantes del científico textil. Algunas gráficas sencillas podrían mostrar la clara distinción entre las muestras. Véase la figura 1.5; las medias y la variabilidad muestrales se describen bien en el diagrama de dispersión. El objetivo de este experimento podría ser simplemente determinar cuáles porcentajes de algodón son verdaderamente distintos de los otros. En otras palabras, como en el caso de los datos con nitrógeno y sin nitrógeno, ¿para cuáles porcentajes de algodón existen diferencias claras entre las poblaciones o, de forma más específica, entre las medias de las poblaciones? En este caso quizás un modelo razonable es que cada muestra proviene de una distribución normal. Aquí el objetivo es muy semejante al de los datos con nitrógeno y sin nitrógeno, excepto que se incluyen más muestras. El formalismo del análisis implica nociones de prueba de hipótesis, los cuales se examinarán en el capítulo 10. A propósito, quizás este formalismo no sea necesario a la luz del diagrama de diagnóstico. Pero, ¿describe éste el objetivo real del experimento y, por consiguiente, el enfoque adecuado para el análisis de datos? Es probable que el científico anticipe la existencia de una *resistencia a la tensión máxima de la media de la población* en el rango de concentración de algodón en el experimento. Aquí el análisis de los datos debería girar en torno a un tipo diferente de modelo, es decir, uno

que postule un tipo de estructura que relacione la resistencia a la tensión de la media de la población con la concentración de algodón. En otras palabras, un modelo se puede escribir como

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2,$$

en donde $\mu_{t,c}$ es la resistencia a la tensión de la media de la población, que varía con la cantidad de algodón en el producto C. La implicación de este modelo es que, para un nivel fijo de algodón, hay una población de mediciones de resistencia a la tensión y la media de la población es $\mu_{t,c}$. Este tipo de modelo, que se denomina **modelo de regresión**, se estudiará en los capítulos 11 y 12. La forma funcional la elige el científico. A veces el análisis de datos puede sugerir que se cambie el modelo. Entonces el analista de datos "considera" un modelo que se pueda alterar después de hacer cierto análisis. El uso de un modelo empírico va acompañado por la **teoría de estimación**, donde β_0, β_1 y β_2 se estiman a partir de los datos. Además, la inferencia estadística se puede, entonces, utilizar para determinar lo adecuado del modelo.

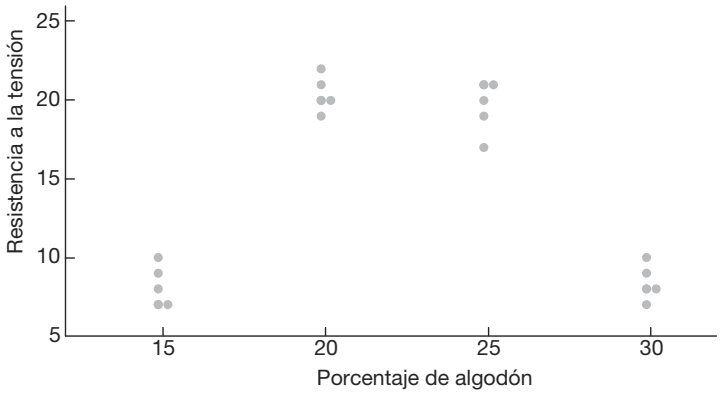


Figura 1.5: Diagrama de dispersión de la resistencia a la tensión y los porcentajes de algodón.

Aquí se hacen evidentes dos puntos de las dos ilustraciones de datos: 1) el tipo de modelo que se emplea para describir los datos a menudo depende del objetivo del experimento, y 2) la estructura del modelo debería aprovechar el insumo científico no estadístico. La selección de un modelo representa una **suposición fundamental** sobre la que se basa la inferencia estadística resultante. A lo largo del libro se hará evidente la importancia que las gráficas pueden llegar a tener. A menudo las gráficas ilustran información que permite que los resultados de la inferencia estadística formal se comuniquen mejor al científico o al ingeniero. A veces las gráficas o el **análisis exploratorio de los datos** pueden enseñar al analista información que no se obtiene del análisis formal. Casi cualquier análisis formal requiere suposiciones que se desarrollan a partir del modelo de datos. Las gráficas pueden resaltar la **violación de suposiciones** que de otra forma no se notarían. A lo largo del libro las gráficas se utilizarán de manera extensa para complementar el análisis formal de los datos. En las siguientes secciones se presentan algunas herramientas gráficas que son útiles para el análisis exploratorio o descriptivo de los datos.

Diagrama de tallo y hojas

Los datos estadísticos obtenidos de poblaciones grandes pueden ser muy útiles para estudiar el comportamiento de la distribución si se presentan en una combinación tabular y gráfica conocida como **diagrama de tallo y hojas**.

Para ejemplificar la elaboración de un diagrama de tallo y hojas considere los datos de la tabla 1.4, que especifican la "vida" de 40 baterías para automóvil similares, registradas al décimo de año más cercano. Las baterías se garantizan por tres años. Comience por dividir cada observación en dos partes: una para el tallo y otra para las hojas, de manera que el tallo represente el dígito entero que antecede al decimal y la hoja corresponda a la parte decimal del número. En otras palabras, para el número 3.7 el dígito 3 se designa al tallo y el 7 a la hoja. Para nuestros datos los cuatro tallos 1, 2, 3 y 4 se listan verticalmente del lado izquierdo de la tabla 1.5, en tanto que las hojas se registran en el lado derecho correspondiente al valor del tallo adecuado. Entonces, la hoja 6 del número 1.6 se registra enfrente del tallo 1; la hoja 5 del número 2.5 enfrente del tallo 2; y así sucesivamente. El número de hojas registrado junto a cada tallo se anota debajo de la columna de frecuencia.

Tabla 1.4: Vida de las baterías para automóvil

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Tabla 1.5: Diagrama de tallo y hojas de la vida de las baterías

Tallo	**Hoja**	**Frecuencia**
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

El diagrama de tallo y hojas de la tabla 1.5 contiene sólo cuatro tallos y, en consecuencia, no ofrece una representación adecuada de la distribución. Para solucionar este problema es necesario aumentar el número de tallos en nuestro diagrama. Una manera sencilla de hacerlo consiste en escribir dos veces cada valor del tallo y después registrar las hojas 0, 1, 2, 3 y 4 enfrente del valor del tallo adecuado, donde aparezca por primera vez; y las hojas 5, 6, 7, 8 y 9 enfrente de este mismo valor del tallo, donde aparece la segunda vez. El diagrama doble de tallo y hojas modificado se ilustra en la tabla 1.6, donde los tallos que corresponden a las hojas 0 a 4 fueron codificados con el símbolo $\star$ y los tallos correspondientes a las hojas 5 a 9 con el símbolo $\bullet$.

En cualquier problema dado debemos decidir cuáles son los valores del tallo adecuados. Esta decisión se toma hasta cierto punto de manera arbitraria, aunque debemos guiarnos por el tamaño de nuestra muestra. Por lo general elegimos entre 5 y 20 tallos. Cuanto más pequeña sea la cantidad de datos disponibles, más pequeña será nuestra elección del número de tallos. Por ejemplo, si los datos constan de números del 1 al 21,

los cuales representan el número de personas en la fila de una cafetería en 40 días laborables seleccionados al azar, y elegimos un diagrama doble de tallo y hojas, los tallos serían 0★, 0•, 1★, 1• y 2★, de manera que la observación de 1 más pequeña tiene tallo 0★ y hoja 1, el número 18 tiene tallo 1• y hoja 8, y la observación de 21 más grande tiene tallo 2★ y hoja 1. Por otro lado, si los datos constan de números de $18,800 a $19,600, que representan las mejores ventas posibles de 100 automóviles nuevos, obtenidos de cierto concesionario, y elegimos un diagrama sencillo de tallo y hojas, los tallos serían 188, 189, 190, …, 196 y las hojas contendrían ahora dos dígitos cada una. Un automóvil que se vende en $19,385 tendría un valor de tallo de 193 y 85 en los dos dígitos de la hoja. En el diagrama de tallo y hojas, las hojas de dígitos múltiples que pertenecen al mismo tallo por lo regular están separadas por comas. En los datos generalmente se ignoran los puntos decimales cuando todos los números a la derecha del punto decimal representan hojas, como en el caso de las tablas 1.5 y 1.6. Sin embargo, si los datos constaran de números que van de 21.8 a 74.9, podríamos elegir los dígitos 2, 3, 4, 5, 6 y 7 como los tallos, de manera que un número como 48.3 tendría un valor de tallo de 4 y un valor de hoja de 8.3.

Tabla 1.6: Diagrama doble de tallo y hojas para la vida de las baterías

Tallo	Hoja	Frecuencia
1·	69	2
2⋆	2	1
2·	5669	4
3⋆	001111222333444	15
3·	5567778899	10
4⋆	11234	5
4·	577	3

El diagrama de tallo y hojas representa una manera eficaz de resumir los datos. Otra forma consiste en el uso de la **distribución de frecuencias**, donde los datos, agrupados en diferentes clases o intervalos, se pueden construir contando las hojas que pertenecen a cada tallo y considerando que cada tallo define un intervalo de clase. En la tabla 1.5 el tallo 1 con 2 hojas define el intervalo 1.0-1.9, que contiene 2 observaciones; el tallo 2 con 5 hojas define el intervalo 2.0-2.9, que contiene 5 observaciones; el tallo 3 con 25 hojas define el intervalo 3.0-3.9, con 25 observaciones; y el tallo 4 con 8 hojas define el intervalo 4.0-4.9, que contiene 8 observaciones. Para el diagrama doble de tallo y hojas de la tabla 1.6 los tallos definen los siete intervalos de clase 1.5-1.9, 2.0-2.4, 2.5-2.9, 3.0-3.4, 3.5-3.9, 4.0-4.4 y 4.5-4.9, con frecuencias 2, 1, 4, 15, 10, 5 y 3, respectivamente.

Histograma

Al dividir cada frecuencia de clase entre el número total de observaciones, obtenemos la proporción del conjunto de observaciones en cada una de las clases. Una tabla que lista las frecuencias relativas se denomina **distribución de frecuencias relativas**. En la tabla 1.7 se presenta la distribución de frecuencias relativas para los datos de la tabla 1.4, que muestra los puntos medios de cada intervalo de clase.

La información que brinda una distribución de frecuencias relativas en forma tabular es más fácil de entender si se presenta en forma gráfica. Con los puntos medios de

Tabla 1.7: Distribución de frecuencias relativas de la vida de las baterías

Intervalo de clase	Punto medio de la clase	Frecuencia, f	Frecuencia relativa
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

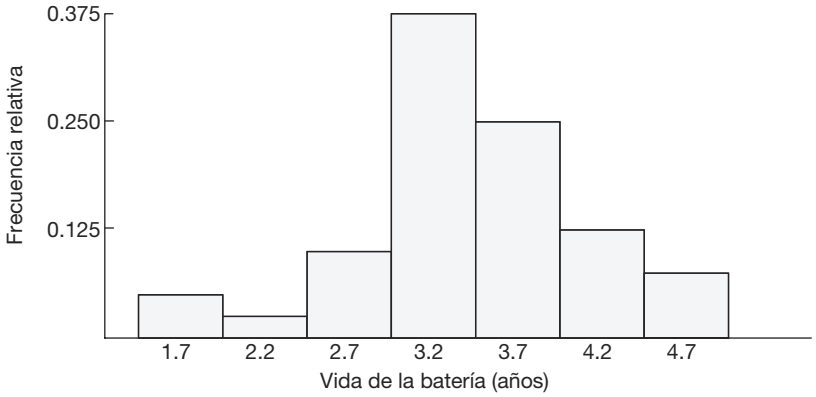


Figura 1.6: Histograma de frecuencias relativas.

cada intervalo y las frecuencias relativas correspondientes construimos un **histograma de frecuencias relativas** (figura 1.6).

Muchas distribuciones de frecuencias continuas se pueden representar gráficamente mediante la curva en forma de campana característica de la figura 1.7. Herramientas gráficas como las de las figuras 1.6 y 1.7 ayudan a comprender la naturaleza de la población. En los capítulos 5 y 6 examinaremos una propiedad de la población que se conoce como **distribución**. Aunque más adelante en este texto se proporcionará una definición más precisa de una distribución o de una **distribución de probabilidad**, aquí podemos visualizarla como la que se podría haber visto en el límite de la figura 1.7 cuando el tamaño de la muestra aumentara.

Se dice que una distribución es **simétrica** si se puede doblar a lo largo de un eje vertical de manera que ambos lados coincidan. Si una distribución carece de simetría respecto de un eje vertical, se dice que está **sesgada**. La distribución que se ilustra en la figura 1.8a se dice que está sesgada a la derecha porque tiene una cola derecha larga y una cola izquierda mucho más corta. En la figura 1.8b observamos que la distribución es simétrica; mientras que en la figura 1.8c está sesgada a la izquierda.

Al girar un diagrama de tallo y hojas en dirección contraria a la de las manecillas del reloj en un ángulo de 90°, vemos que las columnas de hojas que resultan forman una imagen parecida a un histograma. Por lo tanto, si nuestro objetivo principal al observar los datos es determinar la forma general o la forma de la distribución, rara vez será necesario construir un histograma de frecuencias relativas.

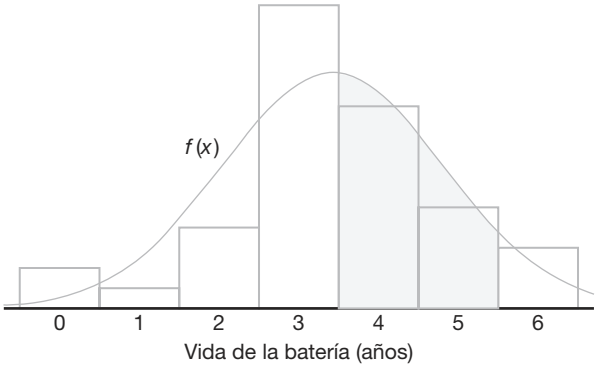


Figura 1.7: Estimación de la distribución de frecuencias.

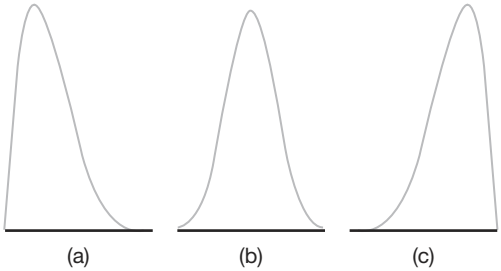


Figura 1.8: Sesgo de los datos.

Gráfica de caja y bigote o gráfica de caja

Otra presentación que es útil para reflejar propiedades de una muestra es la **gráfica de caja y bigote**, la cual encierra el *rango intercuartil* de los datos en una caja que contiene la mediana representada. El rango intercuartil tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior). Además de la caja se prolongan "bigotes", que indican las observaciones alejadas en la muestra. Para muestras razonablemente grandes la presentación indica el centro de localización, la variabilidad y el grado de asimetría.

Además, una variación denominada **gráfica de caja** puede ofrecer al observador información respecto de cuáles observaciones son **valores extremos**. Los valores extremos son observaciones que se consideran inusualmente alejadas de la masa de datos. Existen muchas pruebas estadísticas diseñadas para detectar este tipo de valores. Técnicamente se puede considerar que un valor extremo es una observación que representa un "evento raro" (existe una probabilidad pequeña de obtener un valor que esté lejos de la masa de datos). El concepto de valores extremos volverá a surgir en el capítulo 12 en el contexto del análisis de regresión.

La información visual en las gráficas de caja y bigote o en las de caja no intenta ser una prueba formal de valores extremos, más bien se considera una herramienta de diagnóstico. Aunque la determinación de cuáles observaciones son valores extremos varía de acuerdo con el tipo de software que se emplee, un procedimiento común para determinarlo consiste en utilizar un **múltiplo del rango intercuartil**. Por ejemplo, si la distancia desde la caja excede 1.5 veces el rango intercuartil (en cualquier dirección), la observación se podría considerar un valor extremo.

Ejemplo 1.5: Se midió el contenido de nicotina en una muestra aleatoria de 40 cigarrillos. Los datos se presentan en la tabla 1.8.

Tabla 1.8: Valores de nicotina para el ejemplo 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

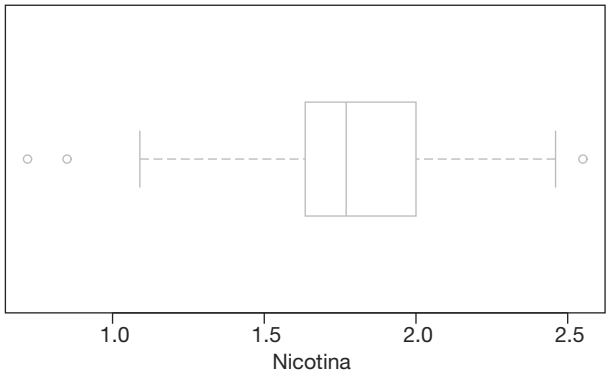


Figura 1.9: Gráfica de caja y bigote para el ejemplo 1.5.

La figura 1.9 muestra la gráfica de caja y bigote de los datos, la cual describe las observaciones 0.72 y 0.85 como valores extremos moderados en la cola inferior; en tanto que la observación 2.55 es un valor extremo moderado en la cola superior. En este ejemplo el rango intercuartil es 0.365, y 1.5 veces el rango intercuartil es 0.5475. Por otro lado, la figura 1.10 presenta un diagrama de tallo y hojas.

Ejemplo 1.6: Considere los datos de la tabla 1.9, que constan de 30 muestras que miden el grosor de las "asas" de latas de pintura (véase el trabajo de Hogg y Ledolter de 1992 en la bibliografía). La figura 1.11 describe una gráfica de caja y bigote para este conjunto asimétrico de datos. Observe que el bloque izquierdo es considerablemente más grande que el bloque de la derecha. La mediana es 35. El cuartil inferior es 31, mientras que el superior es 36. Advierta también que la observación alejada de la derecha está más lejos de la caja que la observación extrema de la izquierda. No hay valores extremos en este conjunto de datos.

```
El punto decimal se encuentra 1 dígito(s) a la izquierda de I
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Figura 1.10: Diagrama de tallo y hojas para los datos de nicotina.

Tabla 1.9: Datos para el ejemplo 1.6

Muestra	Mediciones	Muestra	Mediciones
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

Existen otras formas en las que las gráficas de caja y bigote, y otras presentaciones gráficas, pueden ayudar al analista. Las muestras múltiples se pueden comparar de forma gráfica. Los diagramas de los datos pueden sugerir relaciones entre las variables y las gráficas ayudan a detectar anomalías u observaciones extremas en las muestras.

Existen otros tipos diferentes de diagramas y herramientas gráficas, los cuales se estudiarán en el capítulo 8 después de presentar otros detalles teóricos.

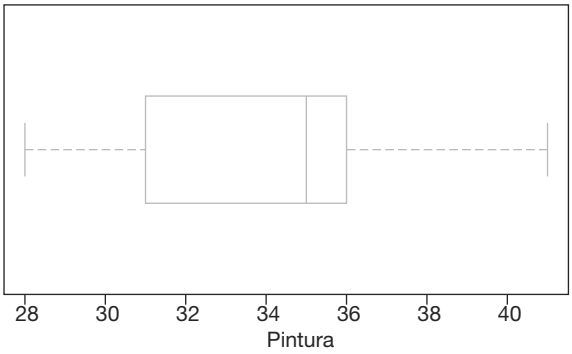


Figura 1.11: Gráfica de caja y bigote del grosor de las "asas" de latas de pintura.

Otras características distintivas de una muestra

Hay características de la distribución o de la muestra, además de las medidas del centro de localización y variabilidad, que definen aún más su naturaleza. Por ejemplo, en tanto que la mediana divide los datos (o su distribución) en dos partes, existen otras medidas que dividen partes o segmentos de la distribución que pueden ser muy útiles. Una separación en cuatro partes se hace mediante *cuartiles*, donde el tercer cuartil separa el cuarto (25%) superior del resto de los datos, el segundo cuartil es la mediana y el primer cuartil separa el cuarto (25%) inferior del resto de los datos. La distribución puede dividirse incluso más detalladamente calculando los percentiles. Tales cantidades dan al analista una noción de las denominadas *colas* de la distribución (es decir, los valores que son relativamente extremos, ya sean pequeños o grandes). Por ejemplo, el percentil 95 separa el 5% superior del 95% inferior. Para los extremos en la parte inferior o *cola inferior* de la distribución prevalecen definiciones similares. El primer percentil separa el 1% inferior del resto de la distribución. El concepto de percentiles desempeñará un papel significativo en buena parte de lo que estudiaremos en los siguientes capítulos.

1.7 Tipos generales de estudios estadísticos: diseño experimental, estudio observacional y estudio retrospectivo

En las siguientes secciones destacaremos el concepto de muestreo de una población y el uso de los métodos estadísticos para aprender o quizá para reafirmar la información relevante acerca de una población. La información que se busca y que se obtiene mediante el uso de tales métodos estadísticos a menudo influye en la toma de decisiones, así como en la resolución de problemas en diversas áreas importantes de ingeniería y científicas. Como ilustración, el ejemplo 1.3 describe un experimento sencillo, en el cual los resultados brindan ayuda para determinar los tipos de condiciones en los que no se recomienda utilizar una aleación de aluminio específica que podría ser muy vulnerable a la corrosión. Los resultados serían útiles no sólo para quienes fabrican la aleación, sino también para los clientes que consideren adquirirla. Este caso, y muchos otros que se incluyen en los capítulos 13 a 15, resaltan el concepto de condiciones experimentales diseñadas o controladas (combinaciones de condiciones de recubrimiento y humedad), que son de interés para aprender sobre algunas características o mediciones (nivel de corrosión) que

surgen de tales condiciones. En las mediciones de la corrosión se emplean métodos estadísticos que utilizan tanto medidas de tendencia central como de variabilidad. Como usted verá más adelante en este texto, tales métodos con frecuencia nos guían hacia un modelo estadístico como el que se examinó en la sección 1.6. En este caso el modelo se puede usar para estimar (o predecir) las medidas de la corrosión como una función de la humedad y el tipo de recubrimiento utilizado. De nuevo, para desarrollar este tipo de modelos es muy útil emplear las estadísticas descriptivas que destacan las medidas de tendencia central y de variabilidad.

La información que se ofrece en el ejemplo 1.3 ilustra de manera adecuada los tipos de preguntas de ingeniería que se plantean y se responden aplicando los métodos estadísticos que se utilizan en un diseño experimental y se presentan en este texto. Tales preguntas son las siguientes:

i. ¿Cuál es la naturaleza del efecto de la humedad relativa sobre la corrosión de la aleación de aluminio dentro del rango de humedad relativa en este experimento?

ii. ¿El recubrimiento químico contra la corrosión reduce los niveles de corrosión y existe alguna manera de cuantificar el efecto?

iii. ¿Hay alguna **interacción** entre el tipo de recubrimiento y la humedad relativa que influya en la corrosión de la aleación? Si es así, ¿cómo se podría interpretar?

¿Qué es interacción?

La importancia de las preguntas **i.** y **ii.** debería quedar clara para el lector, ya que ambas tienen que ver con aspectos importantes tanto para los productores como para los usuarios de la aleación. ¿Pero qué sucede con la pregunta **iii.**? El concepto de *interacción* se estudiará con detalle en los capítulos 14 y 15. Considere la gráfica de la figura 1.3, la cual ejemplifica la detección de la interacción entre dos **factores** en un diseño experimental simple. Observe que las líneas que conectan las medias de la muestra no son paralelas. El **paralelismo** habría indicado que el efecto (visto como un resultado de la pendiente de las líneas) de la humedad relativa es igual, es decir, negativo, tanto en la condición sin recubrimiento como en la condición con recubrimiento químico contra la corrosión. Recuerde que la pendiente negativa implica que la corrosión se vuelve más pronunciada a medida que aumenta la humedad. La ausencia de paralelismo implica una interacción entre el tipo de recubrimiento y la humedad relativa. La línea casi "horizontal" para el recubrimiento contra la corrosión, opuesta a la pendiente más pronunciada para la condición sin recubrimiento, sugiere que *el recubrimiento químico contra la corrosión no sólo es benéfico (observe el desplazamiento entre las líneas), sino que la presencia del recubrimiento revela que el efecto de la humedad es despreciable*. Salta a la vista que todas estas cuestiones son muy importantes para el efecto de los dos factores individuales y para la interpretación de la interacción, si está presente.

Los modelos estadísticos son muy útiles para responder preguntas como las descritas en i, ii y iii, en donde los datos provienen de un diseño experimental. Sin embargo, no siempre se cuenta con el tiempo o los recursos que permiten usar un diseño experimental. Por ejemplo, hay muchos casos en los que las condiciones de interés para el científico o el ingeniero simplemente no se pueden implementar *debido a que es imposible controlar los factores importantes*. En el ejemplo 1.3 la humedad relativa y el tipo de recubrimiento (o la ausencia de éste) son bastante fáciles de controlar. Desde luego, se trata del rasgo distintivo de un diseño experimental. En muchos campos los factores a estudiar no pueden ser controlados por diversas razones. Un control riguroso como el del ejemplo 1.3 permite al analista confiar en que las diferencias encontradas (como en los niveles de

corrosión) se deben a los factores que se pueden controlar. Considere el ejercicio 1.6 de la página 13 como otro ejemplo. En este caso suponga que se eligen 24 especímenes de caucho de silicio y que se asignan 12 a cada uno de los niveles de temperatura de vulcanizado. Las temperaturas se controlan cuidadosamente, por lo que éste es un ejemplo de diseño experimental con **un solo factor**, que es la temperatura de vulcanizado. Se podría suponer que las diferencias encontradas en la media de la resistencia a la tensión son atribuibles a las diferentes temperaturas de vulcanizado.

¿Qué sucede si no se controlan los factores?

Suponga que los factores no se controlan y que *no hay asignación aleatoria* a los tratamientos específicos para las unidades experimentales, y que se necesita obtener información a partir de un conjunto de datos. Como ejemplo considere un estudio donde el interés se centra en la relación entre los niveles de colesterol sanguíneo y la cantidad de sodio medida en la sangre. Durante cierto periodo se revisó el colesterol sanguíneo y el nivel de sodio de un grupo de individuos. En efecto, es posible obtener alguna información útil de tal conjunto de datos. Sin embargo, debería quedar claro que no es posible hacer un control estricto de los niveles de sodio. De manera ideal, los sujetos deberían dividirse aleatoriamente en dos grupos, donde uno fuera el asignado a un nivel alto específico de sodio en la sangre, y el otro a un nivel bajo específico de sodio en la sangre, pero es obvio que esto no es posible. Evidentemente los cambios en los niveles de colesterol se deben a cambios en uno o diversos factores que no se controlaron. Este tipo de estudio, sin control de factores, se denomina **estudio observacional**, el cual la mayoría de las veces implica una situación en que los sujetos se observan a través del tiempo.

Los estudios biológicos y biomédicos a menudo tienen que ser observacionales. Sin embargo, este tipo de estudios no se restringen a dichas áreas. Por ejemplo, considere un estudio diseñado para determinar la influencia de la temperatura ambiental sobre la energía eléctrica que consumen las instalaciones de una planta química. Es evidente que los niveles de la temperatura ambiental no se pueden controlar, por lo tanto, la única manera en que se puede supervisar la estructura de los datos es a partir de los datos de la planta a través del tiempo.

Es importante destacar que una diferencia básica entre un experimento bien diseñado y un estudio observacional es la dificultad para determinar la causa y el efecto verdaderos con este último. Asimismo, las diferencias encontradas en la reacción fundamental (por ejemplo, niveles de corrosión, colesterol sanguíneo, consumo de energía eléctrica en una planta) podrían deberse a otros factores subyacentes que no se controlaron. De manera ideal, en un diseño experimental los *factores perturbadores* serían compensados mediante el proceso de aleatoriedad. En realidad, los cambios en los niveles de colesterol sanguíneo podrían deberse a la ingestión de grasa, a la realización de actividad física, etc. El consumo de energía eléctrica podría estar afectado por la cantidad de bienes producidos o incluso por la pureza de éstos.

Otra desventaja de los estudios observacionales, que a menudo se ignora cuando éstos se comparan con experimentos cuidadosamente diseñados, es que, a diferencia de estos últimos, los observacionales están a merced de circunstancias no controladas, naturales, ambientales o de otros tipos, que repercuten en los niveles de los factores de interés. Por ejemplo, en el estudio biomédico acerca de la influencia de los niveles de sodio en la sangre sobre el colesterol sanguíneo es posible que haya, de hecho, una influencia significativa, pero el conjunto de datos específico usado no involucró la suficiente variación observada en los niveles de sodio debido a la naturaleza de los sujetos elegidos. Desde luego, en un diseño experimental el analista elige y controla los niveles de los factores.

Un tercer tipo de estudio estadístico que podría ser muy útil, pero que tiene notables desventajas cuando se le compara con un diseño experimental, es el **estudio retrospectivo**. Esta clase de estudio emplea estrictamente **datos históricos**, que se obtienen durante un periodo específico. Una ventaja evidente de los datos retrospectivos es el bajo costo de la recopilación de datos. Sin embargo, como se podría esperar, también tiene desventajas claras:

i. La validez y la confiabilidad de los datos históricos a menudo son cuestionables.

ii. Si el tiempo es un aspecto relevante en la estructura de los datos, podría haber datos faltantes.

iii. Podrían existir errores en la recopilación de los datos que no se conocen.

iv. De nuevo, como en el caso de los datos observacionales, no hay control en los rangos de las variables a medir (es decir, en los factores a estudiar). De hecho, las variaciones que se encuentran en los datos históricos a menudo no son significativas para estudios actuales.

En la sección 1.6 se puso cierto énfasis en los modelos de las relaciones entre variables. Presentamos el concepto de análisis de regresión, el cual se estudia en los capítulos 11 y 12, y se considera como una forma del análisis de datos para los diseños experimentales que se examinarán en los capítulos 14 y 15. En la sección 1.6 se utilizó a modo de ejemplo un modelo que relaciona la media poblacional de la resistencia a la tensión de la tela con los porcentajes de algodón, en el cual 20 especímenes de tela representaban las unidades experimentales. En este caso, los datos provienen de un diseño experimental simple, en el que los porcentajes de algodón individuales fueron seleccionados por el científico.

Con frecuencia, tanto los datos observacionales como los retrospectivos se utilizan para observar las relaciones entre variables a través de los procedimientos de construcción de modelos que se estudiarán en los capítulos 11 y 12. Aunque las ventajas de los diseños experimentales se pueden aplicar cuando la finalidad es la construcción de un modelo estadístico, hay muchas áreas en las que no es posible diseñar experimentos, de manera que *habrá que utilizar los datos históricos u observacionales*. Aquí nos referimos al conjunto de datos históricos que se incluye en el ejercicio 12.5 de la página 450. El objetivo es construir un modelo que dé como resultado una ecuación o relación que vincule el consumo mensual de energía eléctrica con la temperatura ambiental promedio, x_1, el número de días en el mes, x_2, la pureza promedio del producto, x_3 y las toneladas de bienes producidos, x_4. Se trata de los datos históricos del año anterior.

Ejercicios

1.13 Un fabricante de componentes electrónicos se interesa en determinar el tiempo de vida de cierto tipo de batería. Una muestra, en horas de vida, es como la siguiente:

123, 116, 122, 110, 175, 126, 125, 111, 118, 117.

a) Calcule la media y la mediana de la muestra.
b) ¿Qué característica en este conjunto de datos es la responsable de la diferencia sustancial entre ambas?

1.14 Un fabricante de neumáticos quiere determinar el diámetro interior de un neumático de cierto grado de calidad. Idealmente el diámetro sería de 570 mm. Los datos son los siguientes:

572, 572, 573, 568, 569, 575, 565, 570.

a) Calcule la media y la mediana de la muestra.
b) Obtenga la varianza, la desviación estándar y el rango de la muestra.
c) Con base en los estadísticos calculados en los incisos *a*) y *b*), ¿qué comentaría acerca de la calidad de los neumáticos?

1.15 Cinco lanzamientos independientes de una moneda tienen como resultado *cinco caras.* Resulta que si la moneda es legal, la probabilidad de este resultado es $(1/2)^5 = 0.03125$. ¿Proporciona esto evidencia sólida

de que la moneda no es legal? Comente y utilice el concepto de valor-*P* que se analizó en la sección 1.1.

1.16 Muestre que las n piezas de información en $\sum_{i=1}^{n}(x_i - \bar{x})^2$ no son independientes; es decir, demuestre que

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

1.17 Se realiza un estudio acerca de los efectos del tabaquismo sobre los patrones de sueño. La medición que se observa es el tiempo, en minutos, que toma quedar dormido. Se obtienen los siguientes datos:

Fumadores:	69.3	56.0	22.1	47.6
	53.2	48.1	52.7	34.4
	60.2	43.8	23.2	13.8
No fumadores:	28.6	25.1	26.4	34.9
	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1
	36.0	37.9	13.9	

a) Calcule la media de la muestra para cada grupo.
b) Calcule la desviación estándar de la muestra para cada grupo.
c) Elabore una gráfica de puntos de los conjuntos de datos A y B en la misma línea.
d) Comente qué clase de efecto parece tener el hecho de fumar sobre el tiempo que se requiere para quedarse dormido.

1.18 Las siguientes puntuaciones representan la calificación en el examen final para un curso de estadística elemental:

23	60	79	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

a) Elabore un diagrama de tallo y hojas para las calificaciones del examen, donde los tallos sean 1, 2, 3,…, 9.
b) Elabore un histograma de frecuencias relativas, trace un estimado de la gráfica de la distribución y analice la asimetría de la distribución.
c) Calcule la media, la mediana y la desviación estándar de la muestra.

1.19 Los siguientes datos representan la duración de vida, en años, medida al entero más cercano, de 30 bombas de combustible similares.

2.0	3.0	0.3	3.3	1.3	0.4
0.2	6.0	5.5	6.5	0.2	2.3
1.5	4.0	5.9	1.8	4.7	0.7
4.5	0.3	1.5	0.5	2.5	5.0
1.0	6.0	5.6	6.0	1.2	0.2

a) Construya un diagrama de tallo y hojas para la vida, en años, de las bombas de combustible, utilizando el dígito a la izquierda del punto decimal como el tallo para cada observación.
b) Determine una distribución de frecuencias relativas.
c) Calcule la media, el rango y la desviación estándar de la muestra.

1.20 Los siguientes datos representan la duración de la vida, en segundos, de 50 moscas de la fruta que se someten a un nuevo aerosol en un experimento de laboratorio controlado.

17	20	10	9	23	13	12	19	18	24
12	14	6	9	13	6	7	10	13	7
16	18	8	3	3	32	9	7	10	11
13	7	18	7	10	4	27	19	16	8
7	10	5	14	15	10	9	6	7	15

a) Elabore un diagrama doble de tallo y hojas para el periodo de vida de las moscas de la fruta usando los tallos 0★, 0•, 1★, 1•, 2★, 2• y 3★ de manera que los tallos codificados con los símbolos ★ y • se asocien, respectivamente, con las hojas 0 a 4 y 5 a 9.
b) Determine una distribución de frecuencias relativas.
c) Construya un histograma de frecuencias relativas.
d) Calcule la mediana.

1.21 La duración de fallas eléctricas, en minutos, se presenta en la siguiente tabla.

22	18	135	15	90	78	69	98	102
83	55	28	121	120	13	22	124	112
70	66	74	89	103	24	21	112	21
40	98	87	132	115	21	28	43	37
50	96	118	158	74	78	83	93	95

a) Calcule la media y la mediana muestrales de las duraciones de la falla eléctrica.
b) Calcule la desviación estándar de las duraciones de la falla eléctrica.

1.22 Los siguientes datos son las mediciones del diámetro de 36 cabezas de remache en centésimos de una pulgada.

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75
6.66	6.66	6.64	6.76	6.73	6.80	6.72	6.76
6.76	6.68	6.66	6.62	6.72	6.76	6.70	6.78
6.76	6.67	6.70	6.72	6.74	6.81	6.79	6.78
6.66	6.76	6.76	6.72				

a) Calcule la media y la desviación estándar de la muestra.
b) Construya un histograma de frecuencias relativas para los datos.

c) Comente si existe o no una indicación clara de que la muestra proviene de una población que tiene una distribución en forma de campana.

1.23 En 20 automóviles elegidos aleatoriamente, se tomaron las emisiones de hidrocarburos en velocidad en vacío, en partes por millón (ppm), para modelos de 1980 y 1990.

Modelos 1980:

141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380

Modelos 1990:

140 160 20 20 223 60 20 95 360 70
220 400 217 58 235 380 200 175 85 65

a) Construya una gráfica de puntos como la de la figura 1.1.
b) Calcule la media de la muestra para los dos años y sobreponga las dos medias en las gráficas.
c) Comente sobre lo que indica la gráfica de puntos respecto de si cambiaron o no las emisiones poblacionales de 1980 a 1990. Utilice el concepto de variabilidad en sus comentarios.

1.24 Los siguientes son datos históricos de los sueldos del personal (dólares por alumno) en 30 escuelas seleccionadas de la región este de Estados Unidos a principios de la década de 1970.

3.79 2.99 2.77 2.91 3.10 1.84 2.52 3.22
2.45 2.14 2.67 2.52 2.71 2.75 3.57 3.85
3.36 2.05 2.89 2.83 3.13 2.44 2.10 3.71
3.14 3.54 2.37 2.68 3.51 3.37

a) Calcule la media y la desviación estándar de la muestra.
b) Utilice los datos para elaborar un histograma de frecuencias relativas.
c) Construya un diagrama de tallo y hojas con los datos.

1.25 El siguiente conjunto de datos se relaciona con el ejercicio 1.24 y representa el porcentaje de las familias que se ubican en el nivel superior de ingresos en las mismas escuelas individuales y con el mismo orden del ejercicio 1.24.

72.2 31.9 26.5 29.1 27.3 8.6 22.3 26.5
20.4 12.8 25.1 19.2 24.1 58.2 68.1 89.2
55.1 9.4 14.5 13.9 20.7 17.9 8.5 55.4
38.1 54.2 21.5 26.2 59.1 43.3

a) Calcule la media de la muestra.
b) Calcule la mediana de la muestra.
c) Construya un histograma de frecuencias relativas con los datos.
d) Determine la media recortada al 10%. Compárela con los resultados de los incisos *a*) y *b*) y exprese su comentario.

1.26 Suponga que le interesa emplear los conjuntos de datos de los ejercicios 1.24 y 1.25 para derivar un modelo que prediga los salarios del personal como una función del porcentaje de familias en un nivel alto de ingresos para los sistemas escolares actuales. Comente sobre cualquier desventaja de llevar a cabo este tipo de análisis.

1.27 Se realizó un estudio para determinar la influencia del desgaste, y, de un cojinete como una función de la carga, x, sobre el cojinete. Para este estudio se utilizó un diseño experimental con tres niveles de carga: 700 lb, 1000 lb y 1300 lb. En cada nivel se utilizaron cuatro especímenes y las medias muestrales fueron 210, 325 y 375, respectivamente.

a) Grafique el promedio de desgaste contra la carga.
b) A partir de la gráfica del inciso *a*), ¿consideraría que hay una relación entre desgaste y carga?
c) Suponga que tenemos los siguientes valores individuales de desgaste para cada uno de los cuatro especímenes en los respectivos niveles de carga. (Vea los datos que siguen). Grafique los resultados de desgaste para todos los especímenes contra los tres valores de carga.
d) A partir de la gráfica del inciso *c*), ¿consideraría que hay una relación clara? Si su respuesta difiere de la del inciso *b*), explique por qué.

	x		
	700	1000	1300
y_1	145	250	150
y_2	105	195	180
y_3	260	375	420
y_4	330	480	750
	$\bar{y}_1 = 210$	$\bar{y}_2 = 325$	$\bar{y}_3 = 375$

1.28 En Estados Unidos y otros países muchas compañías de manufactura utilizan partes moldeadas como componentes de un proceso. La contracción a menudo es un problema importante. Por consiguiente, un dado de metal moldeado para una parte se construye más grande que el tamaño nominal con el fin de permitir su contracción. En un estudio de moldeado por inyección se descubrió que en la contracción influyen múltiples factores, entre los cuales están la velocidad de la inyección en pies/segundo y la temperatura de moldeado en °C. Los dos conjuntos de datos siguientes muestran los resultados del diseño experimental, en donde la velocidad de inyección se mantuvo a dos niveles (bajo y alto) y la temperatura de moldeado se mantuvo constante en un nivel bajo. La contracción se midió en cm $\times 10^4$.

Los valores de contracción a una velocidad de inyección baja fueron:

72.68 72.62 72.58 72.48 73.07
72.55 72.42 72.84 72.58 72.92

Los valores de contracción a una velocidad de inyección alta fueron:

71.62 71.68 71.74 71.48 71.55
71.52 71.71 71.56 71.70 71.50

a) Construya una gráfica de puntos para ambos conjuntos de datos en la misma gráfica. Sobre ésta indique ambas medias de la contracción, tanto para la velocidad de inyección baja como para la velocidad de inyección alta.

b) Con base en los resultados de la gráfica del inciso *a*), y considerando la ubicación de las dos medias y su sentido de variabilidad, ¿cuál es su conclusión respecto del efecto de la velocidad de inyección sobre la contracción a una temperatura de moldeado baja?

1.29 Utilice los datos del ejercicio 1.24 para elaborar una gráfica de caja.

1.30 A continuación se presentan los tiempos de vida, en horas, de 50 lámparas incandescentes, con esmerilado interno, de 40 watts y 110 voltios, los cuales se tomaron de pruebas forzadas de vida:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

Elabore una gráfica de puntos para estos datos.

1.31 Considere la situación del ejercicio 1.28, pero ahora utilice el siguiente conjunto de datos, en el cual la contracción se mide de nuevo a una velocidad de inyección baja y a una velocidad de inyección alta. Sin embargo, esta vez la temperatura de moldeado se aumenta a un nivel "alto" y se mantiene constante.

Los valores de la contracción a una velocidad de inyección baja fueron:

76.20	76.09	75.98	76.15	76.17
75.94	76.12	76.18	76.25	75.82

Los valores de la contracción a una velocidad de inyección alta fueron:

93.25	93.19	92.87	93.29	93.37
92.98	93.47	93.75	93.89	91.62

a) Igual que en el ejercicio 1.28, elabore una gráfica de puntos con ambos conjuntos de datos en la misma gráfica e identifique las dos medias (es decir, la contracción media para la velocidad de inyección baja y para la velocidad de inyección alta).

b) Igual que en el ejercicio 1.28, comente sobre la influencia de la velocidad de inyección en la contracción para la temperatura de moldeado alta. Tome en cuenta la posición de las dos medias y la variabilidad de cada media.

c) Compare su conclusión en el inciso *b*) actual con la del inciso *b*) del ejercicio 1.28, en el cual la temperatura de moldeado se mantuvo a un nivel bajo. ¿Diría que hay interacción entre la velocidad de inyección y la temperatura de moldeado? Explique su respuesta.

1.32 Utilice los resultados de los ejercicios 1.28 y 1.31 para crear una gráfica que ilustre la interacción evidente entre los datos. Use como guía la gráfica de la figura 1.3 del ejemplo 1.3. ¿El tipo de información que se encontró en los ejercicios 1.28 y 1.31 se habría encontrado en un estudio observacional en el que el analista no hubiera tenido control sobre la velocidad de inyección ni sobre la temperatura de moldeado? Explique su respuesta.

1.33 **Proyecto de grupo:** Registre el tamaño de calzado que usa cada estudiante de su grupo. Utilice las medias y las varianzas muestrales, así como los tipos de gráficas que se estudiaron en este capítulo, para resumir cualquier característica que revele una diferencia entre las distribuciones del tamaño del calzado de hombres y mujeres. Haga lo mismo con la estatura de cada estudiante de su grupo.

Capítulo 2

Probabilidad

2.1 Espacio muestral

En el estudio de la estadística tratamos básicamente con la presentación e interpretación de **resultados fortuitos** que ocurren en un estudio planeado o en una investigación científica. Por ejemplo, en Estados Unidos, y con la finalidad de justificar la instalación de un semáforo, se podría registrar el número de accidentes que ocurren mensualmente en la intersección de Driftwood Lane y Royal Oak Drive; en una fábrica se podrían clasificar los artículos que salen de la línea de ensamble como "defectuosos" o "no defectuosos"; en una reacción química se podría revisar el volumen de gas que se libera cuando se varía la concentración de un ácido. Por ello, quienes se dedican a la estadística a menudo manejan datos numéricos que representan conteos o mediciones, o **datos categóricos** que se podrían clasificar de acuerdo con algún criterio.

En este capítulo, al referirnos a cualquier registro de información, ya sea numérico o categórico, utilizaremos el término **observación**. Por consiguiente, los números 2, 0, 1 y 2, que representan el número de accidentes que ocurrieron cada mes, de enero a abril, durante el año pasado en la intersección de Driftwood Lane y Royal Oak Drive, constituyen un conjunto de observaciones. Lo mismo ocurre con los datos categóricos *N, D, N, N* y *D*, que representan los artículos defectuosos o no defectuosos cuando se inspeccionan cinco artículos y se registran como observaciones.

Los estadísticos utilizan la palabra **experimento** para describir cualquier proceso que genere un conjunto de datos. Un ejemplo simple de experimento estadístico es el lanzamiento de una moneda al aire. En tal experimento sólo hay dos resultados posibles: cara o cruz. Otro experimento podría ser el lanzamiento de un misil y la observación de la velocidad a la que se desplaza en tiempos específicos. Las opiniones de los votantes respecto de un nuevo impuesto sobre las ventas también se pueden considerar como observaciones de un experimento. En estadística nos interesan, en particular, las observaciones que se obtienen al repetir varias veces un experimento. En la mayoría de los casos los resultados dependerán del azar, por lo tanto, no se pueden predecir con certeza. Si un químico realizara un análisis varias veces en las mismas condiciones, obtendría diferentes medidas, las cuales indicarían un elemento de probabilidad en el procedimiento experimental. Aun cuando lancemos una moneda al aire repetidas veces, no podemos tener la certeza de que en un lanzamiento determinado obtendremos cara como resultado. Sin embargo, conocemos el conjunto completo de posibilidades para cada lanzamiento.

Dado lo expuesto en la sección 1.7, en la que se revisaron tres tipos de estudios estadísticos y se dieron varios ejemplos de cada uno, ya deberíamos estar familiarizados con el alcance del término experimento. En cada uno de los tres casos, *diseños experimentales*, *estudios observacionales* y *estudios retrospectivos*, el resultado final fue un conjunto

de datos que, por supuesto, está sujeto a la **incertidumbre**. Aunque sólo uno de ellos tiene la palabra *experimento* en su descripción, el proceso de generar los datos o el proceso de observarlos forma parte de un experimento. El estudio de la corrosión expuesto en la sección 1.2 ciertamente implica un experimento en el que los datos son representados por las mediciones de la corrosión. El ejemplo de la sección 1.7, en el que se observó el colesterol y el sodio en la sangre de un conjunto de individuos, representó un estudio observacional (como lo opuesto a un *diseño* experimental) en el que el proceso incluso generó datos y un resultado sujeto a la incertidumbre; por lo tanto, se trata de un experimento. Un tercer ejemplo en la sección 1.7 consistió en un estudio retrospectivo, en el cual se observaron datos históricos sobre el consumo de energía eléctrica por mes y el promedio mensual de la temperatura ambiental. Aun cuando los datos pueden haber estado archivados durante décadas, el proceso se seguirá considerando un experimento.

Definición 2.1: Al conjunto de todos los resultados posibles de un experimento estadístico se le llama **espacio muestral** y se representa con el símbolo S.

A cada resultado en un espacio muestral se le llama **elemento** o **miembro** del espacio muestral, o simplemente **punto muestral**. Si el espacio muestral tiene un número finito de elementos, podemos *listar* los miembros separados por comas y encerrarlos entre llaves. Por consiguiente, el espacio muestral S, de los resultados posibles cuando se lanza una moneda al aire, se puede escribir como

$$S = \{H, T\},$$

en donde H y T corresponden a "caras" y "cruces", respectivamente.

Ejemplo 2.1: Considere el experimento de lanzar un dado. Si nos interesara el número que aparece en la cara superior, el espacio muestral sería

$$S_1 = \{1, 2, 3, 4, 5, 6\}$$

Si sólo estuviéramos interesados en si el número es par o impar, el espacio muestral sería simplemente

$$S_2 = \{\text{par, impar}\}$$

El ejemplo 2.1 ilustra el hecho de que se puede usar más de un espacio muestral para describir los resultados de un experimento. En este caso, S_1 brinda más información que S_2. Si sabemos cuál elemento ocurre en S_1, podremos indicar cuál resultado tiene lugar en S_2; no obstante, saber lo que pasa en S_2 no ayuda mucho a determinar qué elemento ocurre en S_1. En general, lo deseable sería utilizar un espacio muestral que proporcione la mayor información acerca de los resultados del experimento. En algunos experimentos es útil listar los elementos del espacio muestral de forma sistemática utilizando un **diagrama de árbol**.

Ejemplo 2.2: Un experimento consiste en lanzar una moneda y después lanzarla una segunda vez si sale cara. Si en el primer lanzamiento sale cruz, entonces se lanza un dado una vez. Para listar los elementos del espacio muestral que proporciona la mayor información construimos el diagrama de árbol de la figura 2.1. Las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Si empezamos con la rama superior izquierda y nos movemos a la derecha a lo largo de la primera trayectoria, obtenemos el punto muestral *HH,* que indica la posibilidad de que ocurran caras en dos lanzamientos sucesivos de la moneda. De igual manera, el punto muestral *T*3 indica la posibilidad de que la moneda muestre una cruz seguida por un 3 en el lanzamiento del dado. Al seguir todas las trayectorias, vemos que el espacio muestral es

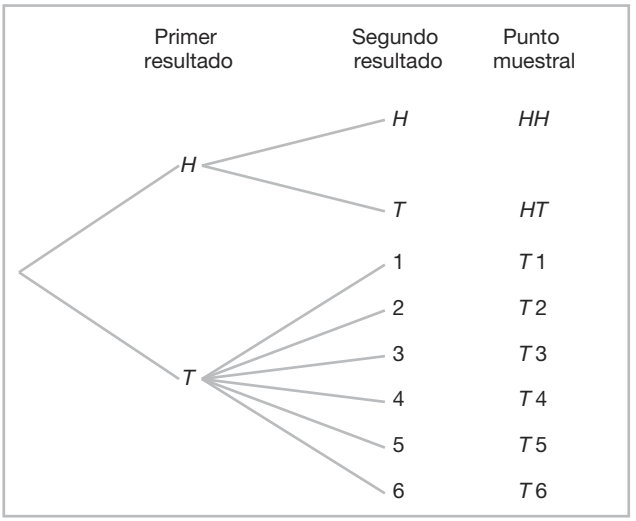


Figura 2.1: Diagrama de árbol para el ejemplo 2.2.

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}.$$

Muchos de los conceptos de este capítulo se ilustran mejor con ejemplos que involucran el uso de dados y cartas. Es particularmente importante utilizar estas aplicaciones al comenzar el proceso de aprendizaje, ya que facilitan el flujo de esos conceptos nuevos en ejemplos científicos y de ingeniería como el siguiente.

Ejemplo 2.3: Suponga que se seleccionan, de forma aleatoria, tres artículos de un proceso de fabricación. Cada artículo se inspecciona y se clasifica como defectuoso, *D,* o no defectuoso, *N*. Para listar los elementos del espacio muestral que brinde la mayor información, construimos el diagrama de árbol de la figura 2.2, de manera que las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Al comenzar con la primera trayectoria, obtenemos el punto muestral *DDD,* que indica la posibilidad de que los tres artículos inspeccionados estén defectuosos. Conforme continuamos a lo largo de las demás trayectorias, vemos que el espacio muestral es

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}.$$

Los espacios muestrales con un número grande o infinito de puntos muestrales se describen mejor mediante un **enunciado** o **método de la regla**. Por ejemplo, si el conjunto de resultados posibles de un experimento fuera el conjunto de ciudades en el mundo con una población de más de un millón de habitantes, nuestro espacio muestral se escribiría como

$$S = \{x \mid x \text{ es una ciudad con una población de más de un millón de habitantes}\},$$

que se lee "*S* es el conjunto de todas las *x,* tales que *x* es una ciudad con una población de más de un millón de habitantes". La barra vertical se lee como "tal que". De manera similar, si *S* es el conjunto de todos los puntos (x, y) sobre los límites o el interior de un círculo de radio 2 con centro en el origen, escribimos la **regla**

$$S = \{(x, y) \mid x^2 + y^2 \leq 4\}.$$

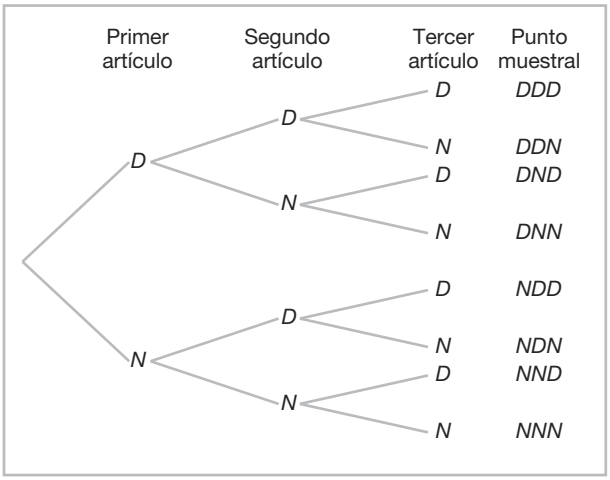


Figura 2.2: Diagrama de árbol para el ejemplo 2.3.

Nuestra elección respecto a describir el espacio muestral utilizando el método de la regla o listando los elementos dependerá del problema específico en cuestión. El método de la regla tiene ventajas prácticas, sobre todo en el caso de muchos experimentos en los que listar se vuelve una tarea tediosa.

Considere la situación del ejemplo 2.3, en el que los artículos que salen del proceso de fabricación están defectuosos, D, o no defectuosos, N. Hay muchos procedimientos estadísticos importantes llamados planes de muestreo, que determinan si un "lote" de artículos se considera o no satisfactorio. Este tipo de planes implican tomar muestras hasta obtener k artículos defectuosos. Suponga que el experimento consiste en tomar muestras de artículos, de forma aleatoria, hasta que salga uno defectuoso. En este caso el espacio muestral sería

$$S = \{D, ND, NND, NNND, \ldots\}.$$

2.2 Eventos

En cualquier experimento dado, podríamos estar interesados en la ocurrencia de ciertos **eventos**, más que en la ocurrencia de un elemento específico en el espacio muestral. Por ejemplo, quizás estemos interesados en el evento A, en el cual el resultado de lanzar un dado es divisible entre 3. Esto ocurrirá si el resultado es un elemento del subconjunto $A = \{3, 6\}$ del espacio muestral S_1 del ejemplo 2.1. Otro ejemplo: podríamos estar interesados en el evento B de que el número de artículos defectuosos sea mayor que 1 en el ejemplo 2.3. Esto ocurrirá si el resultado es un elemento del subconjunto

$$B = \{DDN, DND, NDD, DDD\}$$

del espacio muestral S.

Para cada evento asignamos un conjunto de puntos muestrales, que constituye un subconjunto del espacio muestral. Este subconjunto representa la totalidad de los elementos para los que el evento es cierto.

Definición 2.2: Un **evento** es un subconjunto de un espacio muestral.

Ejemplo 2.4: Dado el espacio muestral $S = \{t \mid t \geq 0\}$, donde t es la vida en años de cierto componente electrónico, el evento A de que el componente falle antes de que finalice el quinto año es el subconjunto $A = \{t \mid 0 \leq t < 5\}$.

Es posible concebir que un evento puede ser un subconjunto que incluye todo el espacio muestral S, o un subconjunto de S que se denomina **conjunto vacío** y se denota con el símbolo ϕ, que no contiene ningún elemento. Por ejemplo, si en un experimento biológico permitimos que A sea el evento de detectar un organismo microscópico a simple vista, entonces $A = \phi$. También, si

$$B = \{x \mid x \text{ es un factor par de } 7\},$$

entonces B debe ser el conjunto vacío, pues los únicos factores posibles de 7 son los números nones 1 y 7.

Considere un experimento en el que se registran los hábitos de tabaquismo de los empleados de una empresa industrial. Un posible espacio muestral podría clasificar a un individuo como no fumador, fumador ocasional, fumador moderado o fumador empedernido. Si se determina que el subconjunto de los fumadores sea un evento, entonces la totalidad de los no fumadores corresponderá a un evento diferente, también subconjunto de S, que se denomina **complemento** del conjunto de fumadores.

Definición 2.3: El **complemento** de un evento A respecto de S es el subconjunto de todos los elementos de S que no están en A. Denotamos el complemento de A mediante el símbolo A'.

Ejemplo 2.5: Sea R el evento de que se seleccione una carta roja de una baraja ordinaria de 52 cartas, y sea S toda la baraja. Entonces R' es el evento de que la carta seleccionada de la baraja no sea una roja sino una negra.

Ejemplo 2.6: Considere el espacio muestral

$$S = \{\text{libro, teléfono celular, mp3, papel, papelería, computadora}\}.$$

Sea $A = \{\text{libro, papelería, computadora, papel}\}$. Entonces, el complemento de A es $A' = \{\text{teléfono celular, mp3}\}$.

Consideremos ahora ciertas operaciones con eventos que darán como resultado la formación de nuevos eventos. Estos eventos nuevos serán subconjuntos del mismo espacio muestral que los eventos dados. Suponga que A y B son dos eventos que se asocian con un experimento. En otras palabras, A y B son subconjuntos del mismo espacio muestral S. Por ejemplo, en el lanzamiento de un dado podríamos hacer que A sea el evento de que ocurra un número par y B el evento de que aparezca un número mayor que 3. Entonces, los subconjuntos $A = \{2, 4, 6\}$ y $B = \{4, 5, 6\}$ son subconjuntos del mismo espacio muestral

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Observe que *tanto* A como B ocurrirán en un lanzamiento dado si el resultado es un elemento del subconjunto $\{4, 6\}$, el cual es precisamente la **intersección** de A y B.

Definición 2.4: La **intersección** de dos eventos A y B, que se denota con el símbolo $A \cap B$, es el evento que contiene todos los elementos que son comunes a A y a B.

Ejemplo 2.7: Sea E el evento de que una persona seleccionada al azar en un salón de clases sea estudiante de ingeniería, y sea F el evento de que la persona sea mujer. Entonces $E \cap F$ es el evento de todas las estudiantes mujeres de ingeniería en el salón de clases.

Ejemplo 2.8: Sean $V = \{a, e, i, o, u\}$ y $C = \{l, r, s, t\}$; entonces, se deduce que $V \cap C = \phi$. Es decir, V y C no tienen elementos comunes, por lo tanto, no pueden ocurrir de forma simultánea. ■

Para ciertos experimentos estadísticos no es nada extraño definir dos eventos, A y B, que no pueden ocurrir de forma simultánea. Se dice entonces que los eventos A y B son **mutuamente excluyentes**. Expresado de manera más formal, tenemos la siguiente definición:

Definición 2.5: Dos eventos A y B son **mutuamente excluyentes** o **disjuntos** si $A \cap B = \phi$; es decir, si A y B no tienen elementos en común.

Ejemplo 2.9: Una empresa de televisión por cable ofrece programas en ocho diferentes canales, tres de los cuales están afiliados con ABC, dos con NBC y uno con CBS. Los otros dos son un canal educativo y el canal de deportes ESPN. Suponga que un individuo que se suscribe a este servicio enciende un televisor sin seleccionar de antemano el canal. Sea A el evento de que el programa pertenezca a la cadena NBC y B el evento de que pertenezca a la cadena CBS. Como un programa de televisión no puede pertenecer a más de una cadena, los eventos A y B no tienen programas en común. Por lo tanto, la intersección $A \cap B$ no contiene programa alguno y, en consecuencia, los eventos A y B son mutuamente excluyentes. ■

A menudo nos interesamos en la ocurrencia de al menos uno de dos eventos asociados con un experimento. Por consiguiente, en el experimento del lanzamiento de un dado, si

$$A = \{2, 4, 6\} \text{ y } B = \{4, 5, 6\},$$

podríamos estar interesados en que ocurran A o B, o en que ocurran tanto A como B. Tal evento, que se llama **unión** de A y B, ocurrirá si el resultado es un elemento del subconjunto $\{2, 4, 5, 6\}$.

Definición 2.6: La **unión** de dos eventos A y B, que se denota con el símbolo $A \cup B$, es el evento que contiene todos los elementos que pertenecen a A o a B, o a ambos.

Ejemplo 2.10: Sea $A = \{a, b, c\}$ y $B = \{b, c, d, e\}$; entonces, $A \cup B = \{a, b, c, d, e\}$. ■

Ejemplo 2.11: Sea P el evento de que un empleado de una empresa petrolera seleccionado al azar fume cigarrillos. Sea Q el evento de que el empleado seleccionado ingiera bebidas alcohólicas. Entonces, el evento $P \cup Q$ es el conjunto de todos los empleados que beben o fuman, o que hacen ambas cosas. ■

Ejemplo 2.12: Si $M = \{x \mid 3 < x < 9\}$ y $N = \{y \mid 5 < y < 12\}$, entonces,

$$M \cup N = \{z \mid 3 < z < 12\}.$$ ■

La relación entre eventos y el correspondiente espacio muestral se puede ilustrar de forma gráfica utilizando **diagramas de Venn**. En un diagrama de Venn representamos el espacio muestral como un rectángulo y los eventos con círculos trazados dentro del rectángulo. De esta forma, en la figura 2.3 vemos que

$$A \cap B = \text{regiones 1 y 2},$$
$$B \cap C = \text{regiones 1 y 3},$$

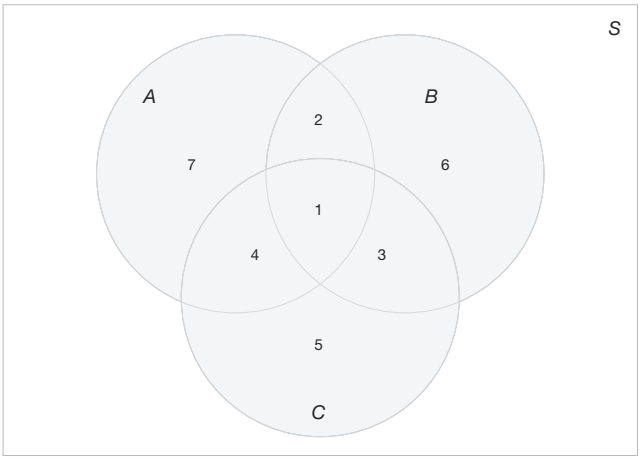


Figura 2.3: Eventos representados por varias regiones.

$$
\begin{aligned}
A \cup C &= \text{regiones 1, 2, 3, 4, 5 y 7,} \\
B' \cap A &= \text{regiones 4 y 7,} \\
A \cap B \cap C &= \text{región 1,} \\
(A \cup B) \cap C' &= \text{regiones 2, 6 y 7,}
\end{aligned}
$$

y así sucesivamente.

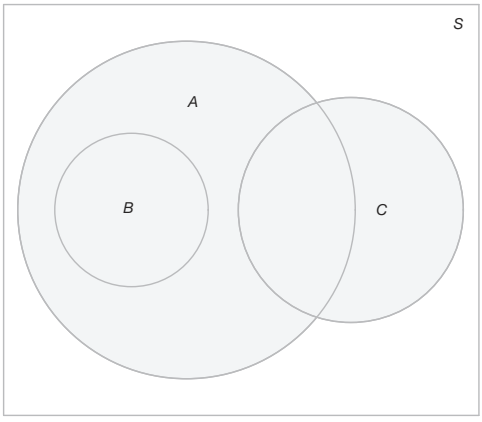


Figura 2.4: Eventos del espacio muestral S.

En la figura 2.4 vemos que los eventos A, B y C son subconjuntos del espacio muestral S. También es claro que el evento B es un subconjunto del evento A; el evento $B \cap C$ no tiene elementos, por lo tanto, B y C son mutuamente excluyentes; el evento $A \cap C$ tiene al menos un elemento; y el evento $A \cup B = A$. Por consiguiente, la figura 2.4 podría representar una situación en la que se selecciona una carta al azar de una baraja ordinaria de 52 cartas y se observa si ocurren los siguientes eventos:

A: la carta es roja,

B: la carta es la jota, la reina o el rey de diamantes,

C: la carta es un as.

Claramente, el evento $A \cap C$ consta sólo de los dos ases rojos.

Varios resultados que se derivan de las definiciones precedentes, y que se pueden verificar de forma sencilla empleando diagramas de Venn, son como los que siguen:

1. $A \cap \phi = \phi$.
2. $A \cup \phi = A$.
3. $A \cap A' = \phi$.
4. $A \cup A' = S$.
5. $S' = \phi$.
6. $\phi' = S$.
7. $(A')' = A$.
8. $(A \cap B)' = A' \cup B'$.
9. $(A \cup B)' = A' \cap B'$.

Ejercicios

2.1 Liste los elementos de cada uno de los siguientes espacios muestrales:

a) el conjunto de números enteros entre 1 y 50 que son divisibles entre 8;
b) el conjunto $S = \{x \mid x^2 + 4x - 5 = 0\}$;
c) el conjunto de resultados cuando se lanza una moneda al aire hasta que aparecen una cruz o tres caras;
d) el conjunto $S = (x \mid x$ es un continente);
e) el conjunto $S = \{x \mid 2x - 4 \geq 0$ y $x < 1\}$.

2.2 Utilice el método de la regla para describir el espacio muestral *S,* que consta de todos los puntos del primer cuadrante dentro de un círculo de radio 3 con centro en el origen.

2.3 ¿Cuáles de los siguientes eventos son iguales?

a) $A = \{1, 3\}$;
b) $B = \{x \mid x$ es un número de un dado$\}$;
c) $C = \{x \mid x^2 - 4x + 3 = 0\}$;
d) $D = \{x \mid x$ es el número de caras cuando se lanzan seis monedas al aire$\}$.

2.4 Un experimento implica lanzar un par de dados, uno verde y uno rojo, y registrar los números que resultan. Si x es igual al resultado en el dado verde y y es el resultado en el dado rojo, describa el espacio muestral S

a) mediante la lista de los elementos (x, y);
b) por medio del método de la regla.

2.5 Un experimento consiste en lanzar un dado y después lanzar una moneda una vez si el número en el dado es par. Si el número en el dado es impar, la moneda se lanza dos veces. Use la notación 4*H,* por ejemplo, para denotar el resultado de que el dado muestre 4 y después la moneda caiga en cara, y 3*HT* para denotar el resultado de que el dado muestre 3, seguido por una cara y después una cruz en la moneda; construya un diagrama de árbol para mostrar los 18 elementos del espacio muestral *S*.

2.6 De un grupo de cuatro suplentes se seleccionan dos jurados para servir en un juicio por homicidio. Utilice la notación A_1A_3, por ejemplo, para denotar el evento simple de que se seleccionen los suplentes 1 y 3, liste los 6 elementos del espacio muestral *S*.

2.7 De un grupo de estudiantes de química se seleccionan cuatro al azar y se clasifican como hombre o mujer. Liste los elementos del espacio muestral S_1 usando la letra *H* para hombre y *M* para mujer. Defina un segundo espacio muestral S_2 donde los elementos representen el número de mujeres seleccionadas.

2.8 Para el espacio muestral del ejercicio 2.4,

a) liste los elementos que corresponden al evento *A* de que la suma sea mayor que 8;
b) liste los elementos que corresponden al evento *B* de que ocurra un 2 en cualquiera de los dos dados;
c) liste los elementos que corresponden al evento *C* de que salga un número mayor que 4 en el dado verde;
d) liste los elementos que corresponden al evento $A \cap C$;
e) liste los elementos que corresponden al evento $A \cap B$;
f) liste los elementos que corresponden al evento $B \cap C$;
g) construya un diagrama de Venn para ilustrar las intersecciones y uniones de los eventos *A, B* y *C*.

2.9 Para el espacio muestral del ejercicio 2.5,

a) liste los elementos que corresponden al evento *A* en el que el dado salga un número menor que 3;
b) liste los elementos que corresponden al evento *B* de que resulten 2 cruces;
c) liste los elementos que corresponden al evento A';

d) liste los elementos que corresponden al evento $A' \cap B$;
e) liste los elementos que corresponden al evento $A \cup B$.

2.10 Se contrata a una empresa de ingenieros para que determine si ciertas vías fluviales en Virginia, Estados Unidos, son seguras para la pesca. Se toman muestras de tres ríos.
a) Liste los elementos de un espacio muestral *S* y utilice las letras *P* para "seguro para la pesca" y *N* para "inseguro para la pesca".
b) Liste los elementos de *S* que correspondan al evento *E* de que al menos dos de los ríos son seguros para la pesca.
c) Defina un evento que tiene como elementos a los puntos

$$\{PPP, NPP, PPN, NPN\}$$

2.11 El currículum de dos aspirantes masculinos para el puesto de profesor de química en una facultad se coloca en el mismo archivo que el de dos aspirantes mujeres. Hay dos puestos disponibles y el primero, con el rango de profesor asistente, se cubre seleccionando al azar a uno de los cuatro aspirantes. El segundo puesto, con el rango de profesor titular, se cubre después mediante la selección aleatoria de uno de los tres aspirantes restantes. Utilice la notación H_2M_1, por ejemplo, para denotar el evento simple de que el primer puesto se cubra con el segundo aspirante hombre y el segundo puesto se cubra después con la primera aspirante mujer,
a) liste los elementos de un espacio muestral *S*;
b) liste los elementos de *S* que corresponden al evento *A* en que el puesto de profesor asistente se cubre con un aspirante hombre;
c) liste los elementos de *S* que corresponden al evento *B* en que exactamente 1 de los 2 puestos se cubre con un aspirante hombre;
d) liste los elementos de *S* que corresponden al evento *C* en que ningún puesto se cubre con un aspirante hombre;
e) liste los elementos de *S* que corresponden al evento $A \cap B$;
f) liste los elementos de *S* que corresponden al evento $A \cup C$;
g) construya un diagrama de Venn para ilustrar las intersecciones y las uniones de los eventos *A, B* y *C*.

2.12 Se estudian el ejercicio y la dieta como posibles sustitutos del medicamento para bajar la presión sanguínea. Se utilizarán tres grupos de individuos para estudiar el efecto del ejercicio. Los integrantes del grupo uno son sedentarios, los del dos caminan y los del tres nadan una hora al día. La mitad de cada uno de los tres grupos de ejercicio tendrá una dieta sin sal. Un grupo adicional de individuos no hará ejercicio ni restringirá su consumo de sal, pero tomará el medicamento estándar. Use *Z* para sedentario, *C* para caminante, *S* para nadador, *Y* para sal, *N* para sin sal, *M* para medicamento y *F* para sin medicamento.
a) Muestre todos los elementos del espacio muestral *S*.
b) Dado que *A* es el conjunto de individuos sin medicamento y *B* es el conjunto de caminantes, liste los elementos de $A \cup B$.
c) Liste los elementos de $A \cap B$.

2.13 Construya un diagrama de Venn para ilustrar las posibles intersecciones y uniones en los siguientes eventos relativos al espacio muestral que consta de todos los automóviles fabricados en Estados Unidos.

C: cuatro puertas, *T:* techo corredizo, *D:* dirección hidráulica

2.14 Si $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ y $A = \{0, 2, 4, 6, 8\}$, $B = \{1, 3, 5, 7, 9\}$, $C = \{2, 3, 4, 5\}$ y $D = \{1, 6, 7\}$, liste los elementos de los conjuntos que corresponden a los siguientes eventos:
a) $A \cup C$;
b) $A \cap B$;
c) C';
d) $(C' \cap D) \cup B$;
e) $(S \cap C)'$;
f) $A \cap C \cap D'$.

2.15 Considere el espacio muestral $S = \{$cobre, sodio, nitrógeno, potasio, uranio, oxígeno, cinc$\}$ y los eventos

$$A = \{\text{cobre, sodio, cinc}\},$$
$$B = \{\text{sodio, nitrógeno, potasio}\}$$
$$C = \{\text{oxígeno}\}.$$

Liste los elementos de los conjuntos que corresponden a los siguientes eventos:
a) A';
b) $A \cup C$;
c) $(A \cap B') \cup C'$;
d) $B' \cap C'$;
e) $A \cap B \cap C$;
f) $(A' \cup B') \cap (A' \cap C)$.

2.16 Si $S = \{x \mid 0 < x < 12\}$, $M = \{x \mid 1 < x < 9\}$ y $N = \{x \mid 0 < x < 5\}$, encuentre
a) $M \cup N$;
b) $M \cap N$;
c) $M' \cap N'$.

2.17 Sean *A, B* y *C* eventos relativos al espacio muestral *S*. Utilice diagramas de Venn para sombrear las áreas que representan los siguientes eventos:
a) $(A \cap B)'$;
b) $(A \cup B)'$;
c) $(A \cap C) \cup B$.

2.18 ¿Cuál de los siguientes pares de eventos son mutuamente excluyentes?

a) Un golfista que se clasifica en último lugar en la vuelta del hoyo 18, en un torneo de 72 hoyos, y pierde el torneo.
b) Un jugador de póquer que tiene flor (todas las cartas del mismo palo) y 3 del mismo palo en la misma mano de 5 cartas.
c) Una madre que da a luz a una niña y a un par de gemelas el mismo día.
d) Un jugador de ajedrez que pierde el último juego y gana el torneo.

2.19 Suponga que una familia sale de vacaciones de verano en su casa rodante y que M es el evento de que sufrirán fallas mecánicas, T es el evento de que recibirán una infracción por cometer una falta de tránsito y V es el evento de que llegarán a un lugar para acampar que esté lleno. Remítase al diagrama de Venn de la figura 2.5 y exprese con palabras los eventos representados por las siguientes regiones:

a) región 5;
b) región 3;
c) regiones 1 y 2 juntas;
d) regiones 4 y 7 juntas;
e) regiones 3, 6, 7 y 8 juntas.

2.20 Remítase al ejercicio 2.19 y al diagrama de Venn de la figura 2.5, liste los números de las regiones que representan los siguientes eventos:

a) La familia no experimentará fallas mecánicas y no será multada por cometer una infracción de tránsito, pero llegará a un lugar para acampar que está lleno.
b) La familia experimentará tanto fallas mecánicas como problemas para localizar un lugar disponible para acampar, pero no será multada por cometer una infracción de tránsito.
c) La familia experimentará fallas mecánicas o encontrará un lugar para acampar lleno, pero no será multada por cometer una infracción de tránsito.
d) La familia no llegará a un lugar para acampar lleno.

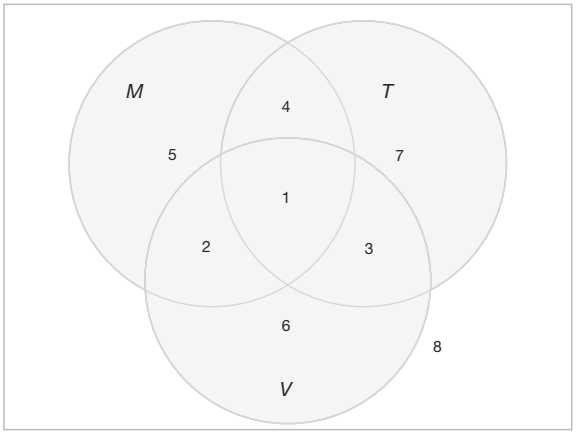


Figura 2.5: Diagrama de Venn para los ejercicios 2.19 y 2.20.

2.3 Conteo de puntos muestrales

Uno de los problemas que el estadístico debe considerar e intentar evaluar es el elemento de aleatoriedad asociado con la ocurrencia de ciertos eventos cuando se realiza un experimento. Estos problemas pertenecen al campo de la probabilidad, un tema que se estudiará en la sección 2.4. En muchos casos debemos ser capaces de resolver un problema de probabilidad mediante el conteo del número de puntos en el espacio muestral, sin listar realmente cada elemento. El principio fundamental del conteo, a menudo denominado **regla de multiplicación**, se establece en la regla 2.1.

Regla 2.1: Si una operación se puede llevar a cabo en n_1 formas, y si para cada una de éstas se puede realizar una segunda operación en n_2 formas, entonces las dos operaciones se pueden ejecutar juntas de $n_1 n_2$ formas.

Ejemplo 2.13: ¿Cuántos puntos muestrales hay en el espacio muestral cuando se lanza un par de dados una vez?

***Solución*:** El primer dado puede caer en cualquiera de $n_1 = 6$ maneras. Para cada una de esas 6 maneras el segundo dado también puede caer en $n_2 = 6$ formas. Por lo tanto, el par de dados puede caer en $n_1 n_2 = (6)(6) = 36$ formas posibles. ■

Ejemplo 2.14: Un urbanista de una nueva subdivisión ofrece a los posibles compradores de una casa elegir entre Tudor, rústica, colonial y tradicional el estilo de la fachada, y entre una planta, dos pisos y desniveles el plano de construcción. ¿En cuántas formas diferentes puede un comprador ordenar una de estas casas?

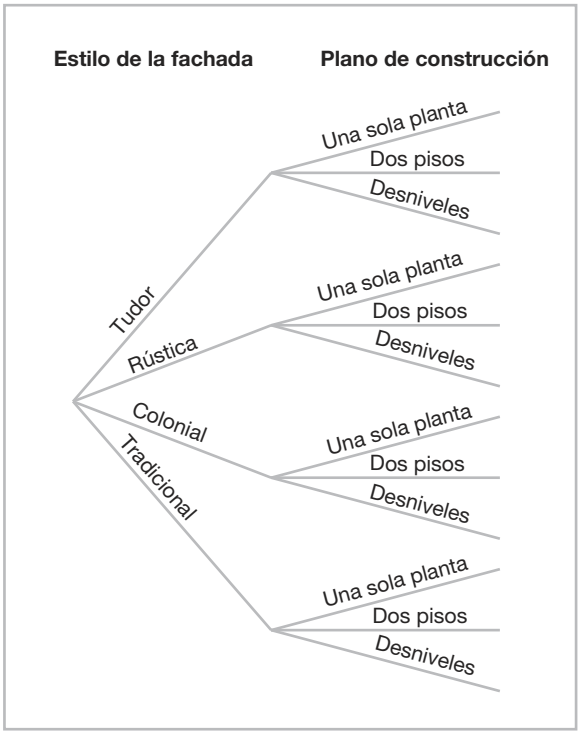


Figura 2.6: Diagrama de árbol para el ejemplo 2.14.

***Solución*:** Como $n_1 = 4$ y $n_2 = 3$, un comprador debe elegir entre

$$n_1 n_2 = (4)(3) = 12 \text{ casas posibles.}$$ ■

Las respuestas a los dos ejemplos anteriores se comprueban construyendo diagramas de árbol y contando las diversas trayectorias a lo largo de las ramas. Así, en el

ejemplo 2.14 habrá $n_1 = 4$ ramas que corresponden a los diferentes estilos de la fachada, y después habrá $n_2 = 3$ ramas que se extienden de cada una de estas 4 ramas para representar los diferentes planos de plantas. Este diagrama de árbol, como se ilustra en la figura 2.6, proporciona las $n_1 n_2 = 12$ opciones de casas dadas por las trayectorias a lo largo de las ramas.

Ejemplo 2.15: Si un miembro de un club que tiene 22 integrantes necesitara elegir un presidente y un tesorero, ¿de cuántas maneras diferentes se podría elegir a ambos?

***Solución*:** Para el puesto de presidente hay 22 posibilidades en total. Para cada una de esas 22 posibilidades hay 21 posibilidades de elegir al tesorero. Si utilizamos la regla de la multiplicación, obtenemos $n_1 \times n_2 = 22 \times 21 = 462$ maneras diferentes. ∎

La regla de la multiplicación (regla 2.1) se puede extender para abarcar cualquier número de operaciones. Por ejemplo, suponga que un cliente desea comprar un nuevo teléfono celular y que puede elegir entre $n_1 = 5$ marcas, $n_2 = 5$ tipos de capacidad y $n_3 = 4$ colores. Estas tres clasificaciones dan como resultado $n_1 n_2 n_3 = (5)(5)(4) = 100$ diferentes formas en las que un cliente puede ordenar uno de estos teléfonos. A continuación se formula la **regla de multiplicación generalizada** que cubre k operaciones.

Regla 2.2: Si una operación se puede ejecutar en n_1 formas, y si para cada una de éstas se puede llevar a cabo una segunda operación en n_2 formas, y para cada una de las primeras dos se puede realizar una tercera operación en n_3 formas, y así sucesivamente, entonces la serie de k operaciones se puede realizar en $n_1 n_2 ... n_k$ formas.

Ejemplo 2.16: Sam va a armar una computadora y para comprar las partes tiene que elegir entre las siguientes opciones: dos marcas de circuitos integrados, cuatro marcas de discos duros, tres marcas de memorias y cinco tiendas locales en las que puede adquirir un conjunto de accesorios. ¿De cuántas formas diferentes puede Sam comprar las partes?

***Solución*:** Como $n_1 = 2$, $n_2 = 4$, $n_3 = 3$ y $n_4 = 5$, hay

$$n_1 \times n_2 \times n_3 \times n_4 = 2 \times 4 \times 3 \times 5 = 120$$

formas diferentes de comprar las partes. ∎

Ejemplo 2.17: ¿Cuántos números pares de cuatro dígitos se pueden formar con los dígitos 0, 1, 2, 5, 6 y 9, si cada dígito se puede usar sólo una vez?

***Solución*:** Como el número debe ser par, tenemos sólo $n_1 = 3$ opciones para la posición de las unidades. Sin embargo, para un número de cuatro dígitos la posición de los millares no puede ser 0. Por lo tanto, consideramos la posición de las unidades en dos partes: 0 o diferente de 0. Si la posición de las unidades es 0 (es decir, $n_1 = 1$), tenemos $n_2 = 5$ opciones para la posición de los millares, $n_3 = 4$ para la posición de las centenas y $n_4 = 3$ para la posición de las decenas. Por lo tanto, en este caso tenemos un total de

$$n_1 n_2 n_3 n_4 = (1)(5)(4)(3) = 60$$

números pares de cuatro dígitos. Por otro lado, si la posición de las unidades no es 0 (es decir, $n_1 = 2$), tenemos $n_2 = 4$ opciones para la posición de los millares, $n_3 = 4$ para la posición de las centenas y $n_4 = 3$ para la posición de las decenas. En esta situación tenemos un total de

$$n_1 n_2 n_3 n_4 = (2)(4)(4)(3) = 96$$

números pares de cuatro dígitos.

Puesto que los dos casos anteriores son mutuamente excluyentes, el número total de números pares de cuatro dígitos se puede calcular usando $60 + 96 = 156$.

Con frecuencia nos interesamos en un espacio muestral que contiene como elementos a todas las posibles ordenaciones o arreglos de un grupo de objetos. Por ejemplo, cuando queremos saber cuántos arreglos diferentes son posibles para sentar a seis personas alrededor de una mesa, o cuando nos preguntamos cuántas ordenaciones diferentes son posibles para sacar dos billetes de lotería de un total de 20. En este caso los diferentes arreglos se llaman **permutaciones**.

Definición 2.7: Una **permutación** es un arreglo de todo o parte de un conjunto de objetos.

Considere las tres letras a, *b* y *c*. Las permutaciones posibles son *abc*, *acb*, *bac*, *bca*, *cab* y *cba,* por lo tanto, vemos que hay 6 arreglos distintos. Si utilizamos la regla 2.2 podemos llegar a la respuesta 6 sin listar realmente las diferentes ordenaciones. Hay $n_1 = 3$ opciones para la primera posición. Sin importar cuál letra se elija, siempre habrá $n_2 = 2$ opciones para la segunda posición. Por último, independientemente de cuál de las dos letras se elija para las primeras dos posiciones, sólo hay $n_3 = 1$ elección para la última posición, lo que da un total de

$$n_1 n_2 n_3 = (3)(2)(1) = 6 \text{ permutaciones}$$

mediante la regla 2.2. En general, *n* objetos distintos se pueden arreglar en

$$n(n-1)(n-2) \ldots (3)(2)(1) \text{ formas.}$$

Existe una notación para una cifra como ésta.

Definición 2.8 Para cualquier entero no negativo *n, n*!, denominado "*n* factorial" se define como

$$N! = n(n-1) \ldots (2)(1),$$

con el caso especial de $0! = 1$.

Si utilizamos el argumento anterior llegamos al siguiente teorema.

Teorema 2.1: El número de permutaciones de *n* objetos es *n*!

El número de permutaciones de las cuatro letras *a*, *b*, *c* y *d* será $4! = 24$. Consideremos ahora el número de permutaciones que son posibles tomando dos de las cuatro letras a la vez. Éstas serían *ab*, *ac*, *ad*, *ba*, *bc*, *bd*, *ca*, *cb*, *cd*, *da*, *db* y *dc*. De nuevo, si utilizamos la regla 2.1, tenemos dos posiciones para llenar con $n_1 = 4$ opciones para la primera y después $n_2 = 3$ opciones para la segunda, para un total de

$$n_1 n_2 = (4)(3) = 12$$

permutaciones. En general, *n* objetos distintos tomados de *r* a la vez se pueden arreglar en

$$n(n-1)(n-2) \ldots (n-r+1)$$

formas. Representamos este producto mediante

$$_nP_r = \frac{n!}{(n-r)!}.$$

Como resultado tenemos el teorema que sigue.

Teorema 2.2: El número de permutaciones de n objetos distintos tomados de r a la vez es

$$_nP_r = \frac{n!}{(n-r)!}.$$

Ejemplo 2.18: En un año se otorgará uno de tres premios (a la investigación, la enseñanza y el servicio) a algunos de los estudiantes, de un grupo de 25, de posgrado del departamento de estadística. Si cada estudiante puede recibir un premio como máximo, ¿cuántas selecciones posibles habría?

***Solución*:** Como los premios son distinguibles, se trata de un problema de permutación. El número total de puntos muestrales es

$$_{25}P_3 = \frac{25!}{(25-3)!} = \frac{25!}{22!} = (25)(24)(23) = 13{,}800.$$

Ejemplo 2.19: En un club estudiantil compuesto por 50 personas se va a elegir a un presidente y a un tesorero. ¿Cuántas opciones diferentes de funcionarios son posibles si

a) no hay restricciones;

b) *A* participará sólo si él es el presidente;

c) *B* y *C* participarán juntos o no lo harán;

d) *D* y *E* no participarán juntos?

***Solución*:** *a*) El número total de opciones de funcionarios, si no hay restricciones, es

$$_{50}P_2 = \frac{50!}{48!} = (50)(49) = 2450.$$

b) Como *A* participaría sólo si es el presidente, tenemos dos situaciones: i) *A* se elige como presidente, lo cual produce 49 resultados posibles para el puesto de tesorero; o ii) los funcionarios se eligen de entre las 49 personas restantes sin tomar en cuenta a A, en cuyo caso el número de opciones es $_{49}P_2 = (49)(48) = 2352$. Por lo tanto, el número total de opciones es $49 + 2352 = 2401$.

c) El número de selecciones cuando *B* y *C* participan juntos es 2. El número de selecciones cuando ni *B* ni *C* se eligen es $_{48}P_2 = 2256$. Por lo tanto, el número total de opciones en esta situación es $2 + 2256 = 2258$.

d) El número de selecciones cuando *D* participa como funcionario pero sin *E* es $(2)(48) = 96$, donde 2 es el número de puestos que *D* puede ocupar y 48 es el número de selecciones de los otros funcionarios de las personas restantes en el club, excepto *E*. El número de selecciones cuando *E* participa como funcionario pero sin *D* también es $(2)(48) = 96$. El número de selecciones cuando tanto *D* como *E* no son elegidos es $_{48}P_2 = 2256$. Por lo tanto, el número total de opciones es $(2)(96) + 2256 = 2448$. Este problema también tiene otra solución rápida: como *D* y *E* sólo pueden participar juntos de dos maneras, la respuesta es $2450 - 2 = 2448$. ∎

Las permutaciones que ocurren al arreglar objetos en un círculo se llaman **permutaciones circulares**. Dos permutaciones circulares no se consideran diferentes a menos que los objetos correspondientes en los dos arreglos estén precedidos o seguidos por un objeto diferente, conforme avancemos en la dirección de las manecillas del reloj. Por ejemplo, si cuatro personas juegan *bridge*, no tenemos una permutación nueva si se mueven una posición en la dirección de las manecillas del reloj. Si consideramos a una persona en una posición fija y arreglamos a las otras tres de 3! formas, encontramos que hay seis arreglos distintos para el juego de *bridge*.

Teorema 2.3: El número de permutaciones de n objetos ordenados en un círculo es $(n-1)!$.

Hasta ahora hemos considerado permutaciones de objetos distintos. Es decir, todos los objetos fueron por completo diferentes o distinguibles. Evidentemente, si tanto la letra b como la c son iguales a x, entonces las 6 permutaciones de las letras a, b y c se convierten en *axx, axx, xax, xax, xxa* y *xxa*, de las cuales sólo 3 son diferentes. Por lo tanto, con 3 letras, en las que 2 son iguales, tenemos $3!/2! = 3$ permutaciones distintas. Con 4 letras diferentes a, b, c y d tenemos 24 permutaciones distintas. Si permitimos que $a = b = x$ y $c = d = y$, podemos listar sólo las siguientes permutaciones distintas: *xxyy, xyxy, yxxy, yyxx, xyyx* y *yxyx*. De esta forma tenemos $4!/(2!2!) = 6$ permutaciones distintas.

Teorema 2.4: El número de permutaciones distintas de n objetos, en el que n_1 son de una clase, n_2 de una segunda clase,..., n_k de una k-ésima clase es

$$\frac{n!}{n_1!n_2!\cdots n_k!}.$$

Ejemplo 2.20: Durante un entrenamiento de fútbol americano colegial, el coordinador defensivo necesita tener a 10 jugadores parados en una fila. Entre estos 10 jugadores hay 1 de primer año, 2 de segundo año, 4 de tercer año y 3 de cuarto año, respectivamente. ¿De cuántas formas diferentes se pueden arreglar en una fila si lo único que los distingue es el grado en el cual están?

***Solución*:** Usando directamente el teorema 2.4, el número total de arreglos es

$$\frac{10!}{1!\,2!\,4!\,3!} = 12{,}600.$$

Con frecuencia nos interesa el número de formas de dividir un conjunto de n objetos en r subconjuntos denominados **celdas**. Se consigue una partición si la intersección de todo par posible de los r subconjuntos es el conjunto vacío ϕ, y si la unión de todos los subconjuntos da el conjunto original. El orden de los elementos dentro de una celda no tiene importancia. Considere el conjunto $\{a, e, i, o, u\}$. Las particiones posibles en dos celdas en las que la primera celda contenga 4 elementos y la segunda 1 son

$$\{(a, e, i, o), (u)\},\ \{(a, i, o, u), (e)\},\ \{(e, i, o, u), (a)\},\ \{(a, e, o, u), (i)\},\ \{(a, e, i, u), (o)\}.$$

Vemos que hay 5 formas de partir un conjunto de 4 elementos en dos subconjuntos o celdas que contengan 4 elementos en la primera celda y 1 en la segunda.

El número de particiones para esta ilustración se denota con la expresión

$$\binom{5}{4,\,1} = \frac{5!}{4!\,1!} = 5,$$

en la que el número superior representa el número total de elementos y los números inferiores representan el número de elementos que van en cada celda. Establecemos esto de forma más general en el teorema 2.5.

Teorema 2.5: El número de formas de partir un conjunto de n objetos en r celdas con n_1 elementos en la primera celda, n_2 elementos en la segunda, y así sucesivamente, es

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!n_2! \cdots n_r!},$$

donde $n_1 + n_2 + \ldots + n_r = n$.

Ejemplo 2.21: Un hotel va a hospedar a siete estudiantes de posgrado que asisten a una conferencia, ¿en cuántas formas los puede asignar a una habitación triple y a dos dobles?

***Solución*:** El número total de particiones posibles sería

$$\binom{7}{3, 2, 2} = \frac{7!}{3!\,2!\,2!} = 210.$$

En muchos problemas nos interesamos en el número de formas de seleccionar r objetos de n sin importar el orden. Tales selecciones se llaman **combinaciones**. Una combinación es realmente una partición con dos celdas, donde una celda contiene los r objetos seleccionados y la otra contiene los $(n - r)$ objetos restantes. El número de tales combinaciones se denota con

$$\binom{n}{r,\, n-r}, \text{ que por lo general se reduce a } \binom{n}{r},$$

debido a que el número de elementos en la segunda celda debe ser $n - r$.

Teorema 2.6: El número de combinaciones de n objetos distintos tomados de r a la vez es

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Ejemplo 2.22: Un niño le pide a su madre que le lleve cinco cartuchos de Game-Boy™ de su colección de 10 juegos recreativos y 5 de deportes. ¿De cuántas maneras podría su madre llevarle 3 juegos recreativos y 2 de deportes?

***Solución*:** El número de formas de seleccionar 3 cartuchos de 10 es

$$\binom{10}{3} = \frac{10!}{3!\,(10-3)!} = 120.$$

El número de formas de seleccionar 2 cartuchos de 5 es

$$\binom{5}{2} = \frac{5!}{2!\,3!} = 10.$$

Si utilizamos la regla de la multiplicación (regla 2.1) con $n_1 = 120$ y $n_2 = 10$, tenemos que hay $(120)(10) = 1200$ formas. ■

Ejemplo 2.23: ¿Cuántos arreglos diferentes de letras se pueden hacer con las letras de la palabra *STATISTICS*?

***Solución*:** Si utilizamos el mismo argumento expuesto en el teorema 2.6, en este ejemplo podemos realmente aplicar el teorema 2.5 para obtener

$$\binom{10}{3, 3, 2, 1, 1} = \frac{10!}{3!\,3!\,2!\,1!\,1!} = 50{,}400.$$

Aquí tenemos 10 letras en total, donde 2 letras (*S*, *T*) aparecen tres veces cada una, la letra *I* aparece dos veces, y las letras *A* y *C* aparecen una vez cada una. Por otro lado, el resultado se puede obtener directamente usando el teorema 2.4. ■

Ejercicios

2.21 A los participantes de una convención se les ofrecen seis recorridos, cada uno de tres días, a sitios de interés. ¿De cuántas maneras se puede acomodar una persona para que vaya a uno de los recorridos planeados por la convención?

2.22 En un estudio médico los pacientes se clasifican en 8 formas de acuerdo con su tipo sanguíneo: AB^+, AB^-, A^+, A^-, B^+, B^-, O^+ u O^-; y también de acuerdo con su presión sanguínea: baja, normal o alta. Encuentre el número de formas en las que se puede clasificar a un paciente.

2.23 Si un experimento consiste en lanzar un dado y después extraer una letra al azar del alfabeto inglés, ¿cuántos puntos habrá en el espacio muestral?

2.24 Los estudiantes de humanidades de una universidad privada se clasifican como estudiantes de primer año, de segundo año, de penúltimo año o de último año, y también de acuerdo con su género (hombres o mujeres). Calcule el número total de clasificaciones posibles para los estudiantes de esa universidad.

2.25 Cierta marca de calzado existe en 5 diferentes estilos y cada estilo está disponible en 4 colores distintos. Si la tienda deseara mostrar la cantidad de pares de zapatos que incluya todos los diversos estilos y colores, ¿cuántos pares diferentes tendría que mostrar?

2.26 Un estudio en California concluyó que siguiendo siete sencillas reglas para la salud un hombre y una mujer pueden prolongar su vida 11 y 7 años en promedio, respectivamente. Estas 7 reglas son: no fumar, hacer ejercicio de manera habitual, moderar su consumo de alcohol, dormir siete u ocho horas, mantener el peso adecuado, desayunar y no ingerir alimentos entre comidas. De cuántas formas puede una persona adoptar cinco de estas reglas:

a) ¿Si la persona actualmente infringe las siete reglas?
b) ¿Si la persona nunca bebe y siempre desayuna?

2.27 Un urbanista de un nuevo fraccionamiento ofrece a un posible comprador de una casa elegir entre 4 diseños, 3 diferentes sistemas de calefacción, un garaje o cobertizo, y un patio o un porche cubierto. ¿De cuántos planos diferentes dispone el comprador?

2.28 Un medicamento para aliviar el asma se puede adquirir en 5 diferentes laboratorios y en forma de líquido, comprimidos o cápsulas, todas en concentración normal o alta. ¿De cuántas formas diferentes puede un médico recetar la medicina a un paciente que sufre de asma?

2.29 En un estudio económico de combustibles, cada uno de 3 autos de carreras se prueba con 5 marcas diferentes de gasolina en 7 lugares de prueba que se localizan en diferentes regiones del país. Si en el estudio se utilizan 2 pilotos y las pruebas se realizan una vez en cada uno de los distintos grupos de condiciones, ¿cuántas pruebas se necesita realizar?

2.30 ¿De cuántas formas distintas se puede responder una prueba de falso-verdadero que consta de 9 preguntas?

2.31 Un testigo de un accidente automovilístico le dijo a la policía que la matrícula del culpable, que huyó, contenía las letras RLH seguidas por 3 dígitos, de los cuales el primero era un 5. Si el testigo no recuerda los 2 últimos dígitos, pero está seguro de que los 3 eran distintos, calcule la cantidad máxima de registros de automóviles que la policía tendría que revisar.

2.32 *a*) ¿De cuántas maneras se pueden formar 6 personas para abordar un autobús?
b) ¿Cuántas maneras son posibles si, de las 6, 3 personas específicas insisten en formarse una después de la otra?
c) ¿De cuántas maneras se pueden formar si, de las 6, 2 personas específicas se rehúsan a formarse una detrás de la otra?

2.33 Si una prueba de opción múltiple consta de 5 preguntas, cada una con 4 respuestas posibles, de las cuales sólo 1 es correcta,
a) ¿de cuántas formas diferentes puede un estudiante elegir una respuesta a cada pregunta?
b) ¿de cuántas maneras puede un estudiante elegir una respuesta a cada pregunta y obtener todas las respuestas incorrectas?

2.34 *a*) ¿Cuántas permutaciones distintas se pueden hacer con las letras de la palabra *COLUMNA*?
b) ¿Cuántas de estas permutaciones comienzan con la letra *M*?

2.35 Un contratista desea construir 9 casas, cada una con diferente diseño. ¿De cuántas formas puede ubicarlas en la calle en la que las va a construir si en un lado de ésta hay 6 lotes y en el lado opuesto hay 3?

2.36 *a*) ¿Cuántos números de tres dígitos se pueden formar con los dígitos 0, 1, 2, 3, 4, 5 y 6 si cada dígito se puede usar sólo una vez?
b) ¿Cuántos de estos números son impares?
c) ¿Cuántos son mayores que 330?

2.37 ¿De cuántas maneras se pueden sentar 4 niños y 5 niñas en una fila, si se deben alternar unos y otras?

2.38 Cuatro parejas compran 8 lugares en la misma fila para un concierto. ¿De cuántas maneras diferentes se pueden sentar…
a) sin restricciones?
b) si cada pareja se sienta junta?
c) si todos los hombres se sientan juntos a la derecha de todas las mujeres?

2.39 En un concurso regional de ortografía, los 8 finalistas son 3 niños y 5 niñas. Encuentre el número de puntos muestrales en el espacio muestral *S* para el número de ordenamientos posibles al final del concurso para
a) los 8 finalistas;
b) los 3 primeros lugares.

2.40 ¿De cuántas formas se pueden cubrir las 5 posiciones iniciales en un equipo de baloncesto con 8 jugadores que pueden jugar cualquiera de las posiciones?

2.41 Encuentre el número de formas en que se puede asignar 6 profesores a 4 secciones de un curso introductorio de psicología, si ningún profesor se asigna a más de una sección.

2.42 De un grupo de 40 boletos se sacan 3 billetes de lotería para el primero, segundo y tercer premios. Encuentre el número de puntos muestrales en *S* para dar los 3 premios, si cada concursante sólo tiene un billete.

2.43 ¿De cuántas maneras se pueden plantar 5 árboles diferentes en un círculo?

2.44 ¿De cuántas formas se puede acomodar en círculo una caravana de ocho carretas de Arizona?

2.45 ¿Cuántas permutaciones distintas se pueden hacer con las letras de la palabra *INFINITO*?

2.46 ¿De cuántas maneras se pueden colocar 3 robles, 4 pinos y 2 arces a lo largo de la línea divisoria de una propiedad, si no se distingue entre árboles del mismo tipo?

2.47 ¿De cuántas formas se puede seleccionar a 3 de 8 candidatos recién graduados, igualmente calificados, para ocupar las vacantes de un despacho de contabilidad?

2.48 ¿Cuántas formas hay en que dos estudiantes no tengan la misma fecha de cumpleaños en un grupo de 60?

2.4 Probabilidad de un evento

Quizá fue la insaciable sed del ser humano por el juego lo que condujo al desarrollo temprano de la teoría de la probabilidad. En un esfuerzo por aumentar sus triunfos, algunos pidieron a los matemáticos que les proporcionaran las estrategias óptimas para los diversos juegos de azar. Algunos de los matemáticos que brindaron tales estrategias fueron Pascal, Leibniz, Fermat y James Bernoulli. Como resultado de este desarrollo inicial de la teoría de la probabilidad, la inferencia estadística, con todas sus predicciones y generalizaciones, ha rebasado el ámbito de los juegos de azar para abarcar muchos otros campos asociados con los eventos aleatorios, como la política, los negocios, el pronóstico del clima y la

investigación científica. Para que estas predicciones y generalizaciones sean razonablemente precisas, resulta esencial la comprensión de la teoría básica de la probabilidad.

¿A qué nos referimos cuando hacemos afirmaciones como "Juan probablemente ganará el torneo de tenis", o "tengo 50% de probabilidades de obtener un número par cuando lanzo un dado", o "la universidad no tiene posibilidades de ganar el juego de fútbol esta noche", o "la mayoría de nuestros graduados probablemente estarán casados dentro de tres años"? En cada caso expresamos un resultado del cual no estamos seguros, pero con base en la experiencia, o a partir de la comprensión de la estructura del experimento, confiamos hasta cierto punto en la validez de nuestra afirmación.

En el resto de este capítulo consideraremos sólo aquellos experimentos para los cuales el espacio muestral contiene un número finito de elementos. La probabilidad de la ocurrencia de un evento que resulta de tal experimento estadístico se evalúa utilizando un conjunto de números reales denominados **pesos** o **probabilidades**, que van de 0 a 1. Para todo punto en el espacio muestral asignamos una probabilidad tal que la suma de todas las probabilidades es 1. Si tenemos razón para creer que al llevar a cabo el experimento es bastante probable que ocurra cierto punto muestral, le tendríamos que asignar a éste una probabilidad cercana a 1. Por el contrario, si creemos que no hay probabilidades de que ocurra cierto punto muestral, le tendríamos que asignar a éste una probabilidad cercana a cero. En muchos experimentos, como lanzar una moneda o un dado, todos los puntos muestrales tienen la misma oportunidad de ocurrencia, por lo tanto, se les asignan probabilidades iguales. A los puntos fuera del espacio muestral, es decir, a los eventos simples que no tienen posibilidades de ocurrir, les asignamos una probabilidad de cero.

Para encontrar la probabilidad de un evento A sumamos todas las probabilidades que se asignan a los puntos muestrales en A. Esta suma se denomina **probabilidad** de A y se denota con $P(A)$.

Definición 2.9: La **probabilidad** de un evento A es la suma de los pesos de todos los puntos muestrales en A. Por lo tanto,

$$0 \le P(A) \le 1, \qquad P(\phi) = 0 \quad \text{y} \quad P(S) = 1.$$

Además, si $A_1, A_2, A_3, \ldots$ es una serie de eventos mutuamente excluyentes, entonces

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots.$$

Ejemplo 2.24 Una moneda se lanza dos veces. ¿Cuál es la probabilidad de que ocurra al menos una cara (H)?

***Solución*:** El espacio muestral para este experimento es

$$S = \{HH, HT, TH, TT\}$$

Si la moneda está balanceada, cada uno de estos resultados tendrá las mismas probabilidades de ocurrir. Por lo tanto, asignamos una probabilidad de ω a cada uno de los puntos muestrales. Entonces, $4\omega = 1$ o $\omega = 1/4$. Si A representa el evento de que ocurra al menos una cara (H), entonces

$$A = \{HH, HT, TH\} \text{ y } P(A) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}.$$

Ejemplo 2.25: Se carga un dado de forma que exista el doble de probabilidades de que salga un número par que uno impar. Si E es el evento de que ocurra un número menor que 4 en un solo lanzamiento del dado, calcule $P(E)$.

***Solución*:** El espacio muestral es $S = \{1, 2, 3, 4, 5, 6\}$. Asignamos una probabilidad de w *a* cada número impar y una probabilidad de $2w$ a cada número par. Como la suma de las probabilidades debe ser 1, tenemos $9w = 1$ o $w = 1/9$. Por lo tanto, asignamos probabilidades de 1/9 y 2/9 a cada número impar y par, respectivamente. Por consiguiente,

$$E = \{1, 2, 3\} \text{ y } P(E) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}.$$

Ejemplo 2.26: En el ejemplo 2.25, sea A el evento de que resulte un número par y sea B el evento de que resulte un número divisible entre 3. Calcule $P(A \cup B)$ y $P(A \cap B)$.

***Solución*:** Para los eventos $A = \{2, 4, 6\}$ y $B = \{3, 6\}$, tenemos

$$A \cup B = \{2, 3, 4, 6\} \text{ y } A \cap B = \{6\}.$$

Al asignar una probabilidad de 1/9 a cada número impar y de 2/9 a cada número par, tenemos

$$P(A \cup B) = \frac{2}{9} + \frac{1}{9} + \frac{2}{9} + \frac{2}{9} = \frac{7}{9} \text{ y } P(A \cap B) = \frac{2}{9}.$$

Si el espacio muestral para un experimento contiene N elementos, todos los cuales tienen las mismas probabilidades de ocurrir, asignamos una probabilidad igual a $1/N$ a cada uno de los N puntos. La probabilidad de que cualquier evento A contenga n de estos N puntos muestrales es entonces el cociente del número de elementos en A y el número de elementos en S.

Regla 2.3: Si un experimento puede dar como resultado cualquiera de N diferentes resultados que tienen las mismas probabilidades de ocurrir, y si exactamente n de estos resultados corresponden al evento A, entonces la probabilidad del evento A es

$$P(A) = \frac{n}{N}.$$

Ejemplo 2.27: A una clase de estadística para ingenieros asisten 25 estudiantes de ingeniería industrial, 10 de ingeniería mecánica, 10 de ingeniería eléctrica y 8 de ingeniería civil. Si el profesor elige al azar a un estudiante para que conteste una pregunta, ¿qué probabilidades hay de que el elegido sea *a*) estudiante de ingeniería industrial, *b*) estudiante de ingeniería civil o estudiante de ingeniería eléctrica?.

***Solución*:** Las especialidades de los estudiantes de ingeniería industrial, mecánica, eléctrica y civil se denotan con I, M, E y C, respectivamente. El grupo está integrado por 53 estudiantes y todos tienen las mismas probabilidades de ser seleccionados.

a) Como 25 de los 53 individuos estudian ingeniería industrial, la probabilidad del evento I, es decir, la de elegir al azar a alguien que estudia ingeniería industrial, es

$$P(I) = \frac{25}{53}.$$

b) Como 18 de los 53 estudiantes son de las especialidades de ingeniería civil o eléctrica, se deduce que

$$P(C \cup E) = \frac{18}{53}.$$

Ejemplo 2.28: En una mano de póquer que consta de 5 cartas encuentre la probabilidad de tener 2 ases y 3 jotas.

***Solución*:** El número de formas de tener 2 ases de 4 cartas es

$$\binom{4}{2} = \frac{4!}{2!\,2!} = 6,$$

y el número de formas de tener 3 jotas de 4 cartas es

$$\binom{4}{3} = \frac{4!}{3!\,1!} = 4.$$

Mediante la regla de multiplicación (regla 2.1), obtenemos $n = (6)(4) = 24$ manos con 2 ases y 3 jotas. El número total de manos de póquer de 5 cartas, todas las cuales tienen las mismas probabilidades de ocurrir, es

$$N = \binom{52}{5} = \frac{52!}{5!\,47!} = 2{,}598{,}960.$$

Por lo tanto, la probabilidad del evento C de obtener 2 ases y 3 jotas en una mano de póquer de 5 cartas es

$$P(C) = \frac{24}{2{,}598{,}960} = 0.9 \times 10^{-5}.$$

Si los resultados de un experimento no tienen las mismas probabilidades de ocurrir, las probabilidades se deben asignar con base en el conocimiento previo o en la evidencia experimental. Por ejemplo, si una moneda no está balanceada, podemos estimar las probabilidades de caras y cruces lanzándola muchas veces y registrando los resultados. De acuerdo con la definición de **frecuencia relativa** de la probabilidad, las probabilidades verdaderas serían las fracciones de caras y cruces que ocurren a largo plazo. Otra forma intuitiva de comprender la probabilidad es el método de la **indiferencia**. Por ejemplo, si usted tiene un dado que cree que está balanceado, el método con el que podría determinar que hay 1/6 de probabilidades de que resulte cada una de las seis caras después de lanzarlo una vez es el método de la indiferencia.

Para encontrar un valor numérico que represente de forma adecuada la probabilidad de ganar en el tenis, dependemos de nuestro desempeño previo en el juego, así como también del de nuestro oponente y, hasta cierto punto, de la capacidad de ganar que creemos tener. De manera similar, para calcular la probabilidad de que un caballo gane una carrera, debemos llegar a una probabilidad basada en las marcas anteriores de todos los caballos que participan en la carrera, así como de las marcas de los jinetes que los montan. La intuición, sin duda, también participa en la determinación del monto que estemos dispuestos a apostar. El uso de la intuición, las creencias personales y otra información indirecta para llegar a probabilidades se conoce como la definición **subjetiva** de la probabilidad.

En la mayoría de las aplicaciones de probabilidad de este libro la que opera es la interpretación de frecuencia relativa de probabilidad, la cual se basa en el experimento estadístico en vez de en la subjetividad y es considerada, más bien, como **frecuencia relativa limitante**. Como resultado, muchas aplicaciones de probabilidad en ciencia e ingeniería se deben basar en experimentos que se puedan repetir. Cuando asignamos probabilidades que se basan en información y opiniones previas, como en la afirmación: "hay grandes probabilidades de que los Gigantes pierdan el Súper Tazón", se encuentran

conceptos menos objetivos de probabilidad. Cuando las opiniones y la información previa difieren de un individuo a otro, la probabilidad subjetiva se vuelve el recurso pertinente. En la estadística bayesiana (véase el capítulo 18) se usará una interpretación más subjetiva de la probabilidad, la cual se basará en obtener información previa de probabilidad.

2.5 Reglas aditivas

A menudo resulta más sencillo calcular la probabilidad de algún evento a partir de las probabilidades conocidas de otros eventos. Esto puede ser cierto si el evento en cuestión se puede representar como la unión de otros dos eventos o como el complemento de algún evento. A continuación se presentan varias leyes importantes que con frecuencia simplifican el cálculo de las probabilidades. La primera, que se denomina **regla aditiva**, se aplica a uniones de eventos.

Teorema 2.7: Si A y B son dos eventos, entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

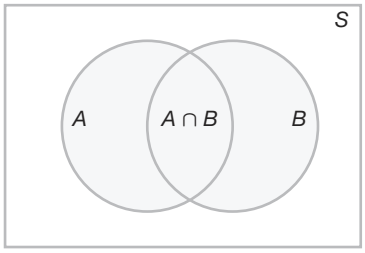


Figura 2.7: Regla aditiva de probabilidad.

***Prueba*:** Considere el diagrama de Venn de la figura 2.7. $P(\underline{A} \cup \underline{B})$ es la suma de las probabilidades de los puntos muestrales en $(A \cup B)$. Así, $P(A) + P(B)$ es la suma de todas las probabilidades en A más la suma de todas las probabilidades en B. Por lo tanto, sumamos dos veces las probabilidades en $(A \cap B)$. Como estas probabilidades se suman a $P(A \cap B)$, debemos restar esta probabilidad una vez para obtener la suma de las probabilidades en $A \cup B$. ∎

Corolario 2.1: Si A y B son mutuamente excluyentes, entonces

$$P(A \cup B) = P(A) + P(B).$$

El corolario 2.1 es un resultado inmediato del teorema 2.7, pues si A y B son mutuamente excluyentes, $A \cap B = 0$ y entonces $P(A \cap B) = P(\phi) = 0$. En general, podemos anotar el corolario 2.2.

Corolario 2.2: Si $A_1, A_2, ..., A_n$ son mutuamente excluyentes, entonces

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n).$$

Un conjunto de eventos $\{A_1, A_2, \ldots A_n\}$ de un espacio muestral S se denomina **partición** de S si $A_1, A_2, \ldots, A_n$ son mutuamente excluyentes y $A_1 \cup A_2 \cup \cdots \cup A_n = S$. Por lo tanto, tenemos

Corolario 2.3: Si $A_1, A_2, ..., A_n$ es una partición de un espacio muestral S, entonces

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = P(S) = 1.$$

Como se esperaría, el teorema 2.7 se extiende de forma análoga.

Teorema 2.8: Para tres eventos A, B y C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Ejemplo 2.29: Al final del semestre John se va a graduar en la facultad de ingeniería industrial de una universidad. Después de tener entrevistas en dos empresas en donde quiere trabajar, determina que la probabilidad que tiene de lograr una oferta de empleo en la empresa A es 0.8, y que la probabilidad de obtenerla en la empresa B es 0.6. Si, por otro lado, considera que la probabilidad de recibir ofertas de ambas empresas es 0.5, ¿qué probabilidad tiene de obtener al menos una oferta de esas dos empresas?

***Solución*:** Si usamos la regla aditiva tenemos

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.8 + 0.6 - 0.5 = 0.9.$$

Ejemplo 2.30: ¿Cuál es la probabilidad de obtener un total de 7 u 11 cuando se lanza un par de dados?

***Solución*:** Sea A el evento de que resulte 7 y B el evento de que salga 11. Ahora bien, para 6 de los 36 puntos muestrales ocurre un total de 7 y sólo para 2 de ellos ocurre un total de 11. Como todos los puntos muestrales tienen la misma probabilidad, tenemos $P(A) = 1/6$ y $P(B) = 1/18$. Los eventos A y B son mutuamente excluyentes, ya que un total de 7 y uno de 11 no pueden ocurrir en el mismo lanzamiento. Por lo tanto,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}.$$

Este resultado también se podría obtener contando el número total de puntos para el evento $A \cup B$, es decir, 8 y escribir

$$P(A \cup B) = \frac{n}{N} = \frac{8}{36} = \frac{2}{9}.$$

El teorema 2.7 y sus tres corolarios deberían ayudar al lector a comprender mejor la probabilidad y su interpretación. Los corolarios 2.1 y 2.2 sugieren el resultado muy intuitivo tratando con la probabilidad de que ocurra al menos uno de varios eventos, sin que puedan ocurrir dos de ellos simultáneamente. La probabilidad de que al menos ocurra uno es la suma de las probabilidades de ocurrencia de los eventos individuales. El tercer corolario simplemente establece que el valor mayor de una probabilidad (unidad) se asigna a todo el espacio muestral *S*.

Ejemplo 2.31: Las probabilidades de que un individuo que compra un automóvil nuevo elija uno de color verde, uno blanco, uno rojo o uno azul son 0.09, 0.15, 0.21 y 0.23, respectivamente, ¿cuál es la probabilidad de que un comprador dado adquiera un automóvil nuevo que tenga uno de esos colores?

***Solución*:** Sean *V*, *B, R* y *A* los eventos de que un comprador seleccione, respectivamente, un automóvil verde, blanco, rojo o azul. Como estos cuatro eventos son mutuamente excluyentes, la probabilidad es

$$\begin{aligned} P(V \cup B \cup R \cup A) &= P(V) + P(B) + P(R) + P(A) \\ &= 0.09 + 0.15 + 0.21 + 0.23 = 0.68. \end{aligned}$$

A menudo es más difícil calcular la probabilidad de que ocurra un evento que calcular la probabilidad de que el evento no ocurra. Si éste es el caso para algún evento *A,* simplemente calculamos primero $P(A')$ y, después, mediante el teorema 2.7, calculamos $P(A)$ por sustracción.

Teorema 2.9: Si A y A' son eventos complementarios, entonces

$$P(A) + P(A') = 1$$

***Prueba*:** Como $A \cup A' = S,$ y los conjuntos A y A' son disjuntos, entonces

$$1 = P(S) = P(A \cup A') = P(A) + P(A')$$

Ejemplo 2.32: Si las probabilidades de que un mecánico automotriz dé servicio a 3, 4, 5, 6, 7, 8 o más vehículos en un día de trabajo dado son 0.12, 0.19, 0.28, 0.24, 0.10 y 0.07, respectivamente, ¿cuál es la probabilidad de que dé servicio al menos a 5 vehículos el siguiente día de trabajo?

***Solución*:** Sea E el evento de que al menos 5 automóviles reciban servicio. Ahora bien, $P(E) = 1 - P(E'),$ donde E' es el evento de que menos de 5 automóviles reciban servicio. Como

$$P(E') = 0.12 + 0.19 = 0.31.$$

del teorema 2.9 se deduce que

$$P(E) = 1 - 0.31 = 0.69.$$

Ejemplo 2.33: Suponga que las especificaciones del fabricante para la longitud del cable de cierto tipo de computadora son 2000 ± 10 milímetros. En esta industria se sabe que el cable pequeño tiene la misma probabilidad de salir defectuoso (de no cumplir con las especificaciones) que el cable grande. Es decir, la probabilidad de que aleatoriamente se produzca un

cable con una longitud mayor que 2010 milímetros es igual a la probabilidad de producirlo con una longitud menor que 1990 milímetros. Se sabe que la probabilidad de que el procedimiento de producción cumpla con las especificaciones es 0.99.

a) ¿Cuál es la probabilidad de que un cable elegido al azar sea muy largo?

b) ¿Cuál es la probabilidad de que un cable elegido al azar sea más grande que 1990 milímetros?

***Solución*:** Sea E el evento de que un cable cumpla con las especificaciones. Sean P y G los eventos de que el cable sea muy pequeño o muy grande, respectivamente. Entonces,

a) $P(E) = 0.99$ y $P(P) = P(G) = (1 - 0.99)/2 = 0.005$.

b) Si la longitud de un cable seleccionado al azar se denota con X, tenemos

$$P(1990 \leq X \leq 2010) = P(E) = 0.99.$$

Como $P(X \geq 2010) = P(G) = 0.005$,

$$P(X \geq 1990) = P(E) + P(G) = 0.995$$

Esto también se resuelve utilizando el teorema 2.9:

$$P(X \geq 1990) + P(X < 1990) = 1.$$

Así, $P(X \geq 1990) = 1 - P(P) = 1 - 0.005 = 0.995$.

Ejercicios

2.49 Encuentre los errores en cada una de las siguientes aseveraciones:

a) Las probabilidades de que un vendedor de automóviles venda 0, 1, 2 o 3 unidades en un día dado de febrero son 0.19, 0.38, 0.29 y 0.15, respectivamente.

b) La probabilidad de que llueva mañana es 0.40 y la probabilidad de que no llueva es 0.52.

c) Las probabilidades de que una impresora cometa 0, 1, 2, 3 o 4 o más errores al imprimir un documento son 0.19, 0.34, −0.25, 0.43 y 0.29, respectivamente.

d) Al sacar una carta de una baraja en un solo intento la probabilidad de seleccionar un corazón es 1/4, la probabilidad de seleccionar una carta negra es 1/2, y la probabilidad de seleccionar una carta de corazones y negra es 1/8.

2.50 Suponga que todos los elementos de S en el ejercicio 2.8 de la página 42 tienen la misma probabilidad de ocurrencia y calcule

a) la probabilidad del evento A;
b) la probabilidad del evento C;
c) la probabilidad del evento $A \cap C$.

2.51 Una caja contiene 500 sobres, de los cuales 75 contienen \$100 en efectivo, 150 contienen \$25 y 275 contienen \$10. Se puede comprar un sobre en \$25. ¿Cuál es el espacio muestral para las diferentes cantidades de dinero? Asigne probabilidades a los puntos muestrales y después calcule la probabilidad de que el primer sobre que se compre contenga menos de \$100.

2.52 Suponga que se descubre que, en un grupo de 500 estudiantes universitarios de último año, 210 fuman, 258 consumen bebidas alcohólicas, 216 comen entre comidas, 122 fuman y consumen bebidas alcohólicas, 83 comen entre comidas y consumen bebidas alcohólicas, 97 fuman y comen entre comidas y 52 tienen esos tres hábitos nocivos para la salud. Si se selecciona al azar a un miembro de este grupo, calcule la probabilidad de que el estudiante

a) fume pero no consuma bebidas alcohólicas;
b) coma entre comidas y consuma bebidas alcohólicas pero no fume;
c) no fume ni coma entre comidas.

2.53 La probabilidad de que una industria estadounidense se ubique en Shanghái, China, es 0.7, la probabilidad de que se ubique en Beijing, China, es 0.4 y la

probabilidad de que se ubique en Shamghái o Beijing, o en ambas ciudades, es 0.8. ¿Cuál es la probabilidad de que la industria se ubique…

a) en ambas ciudades?
b) en ninguna de esas ciudades?

2.54 Basado en su experiencia, un agente bursátil considera que en las condiciones económicas actuales la probabilidad de que un cliente invierta en bonos libres de impuestos es 0.6, la de que invierta en fondos comunes de inversión es 0.3 y la de que invierta en ambos es 0.15. En esta ocasión encuentre la probabilidad de que un cliente invierta

a) en bonos libres de impuestos o en fondos comunes de inversión;
b) en ninguno de esos dos instrumentos.

2.55 Si cada artículo codificado en un catálogo empieza con 3 letras distintas seguidas por 4 dígitos distintos de cero, calcule la probabilidad de seleccionar aleatoriamente uno de estos artículos codificados que tenga como primera letra una vocal y el último dígito sea par.

2.56 Un fabricante de automóviles está preocupado por el posible retiro de su sedán de cuatro puertas con mayor venta. Si fuera retirado habría 0.25 de probabilidad de que haya un defecto en el sistema de frenos, 0.18 de que haya un defecto en la transmisión, 0.17 de que esté en el sistema de combustible y 0.40 de que esté en alguna otra área.

a) ¿Cuál es la probabilidad de que el defecto esté en los frenos o en el sistema de combustible, si la probabilidad de que haya defectos en ambos sistemas de manera simultánea es 0.15?
b) ¿Cuál es la probabilidad de que no haya defecto en los frenos o en el sistema de combustible?

2.57 Si se elige al azar una letra del alfabeto inglés, encuentre la probabilidad de que la letra

a) sea una vocal excepto *y*;
b) esté listada en algún lugar antes de la letra *j*;
c) esté listada en algún lugar después de la letra *g*.

2.58 Se lanza un par de dados. Calcule la probabilidad de obtener

a) un total de 8;
b) máximo un total de 5.

2.59 En una mano de póquer que consta de 5 cartas, encuentre la probabilidad de tener

a) 3 ases;
b) 4 cartas de corazones y 1 de tréboles.

2.60 Si se toman 3 libros al azar, de un librero que contiene 5 novelas, 3 libros de poemas y 1 diccionario, ¿cuál es la probabilidad de que…

a) se seleccione el diccionario?
b) se seleccionen 2 novelas y 1 libro de poemas?

2.61 En un grupo de 100 estudiantes graduados de preparatoria, 54 estudiaron matemáticas, 69 estudiaron historia y 35 cursaron matemáticas e historia. Si se selecciona al azar uno de estos estudiantes, calcule la probabilidad de que

a) el estudiante haya cursado matemáticas o historia;
b) el estudiante no haya llevado ninguna de estas materias;
c) el estudiante haya cursado historia pero no matemáticas.

2.62 La empresa Dom's Pizza utiliza pruebas de sabor y el análisis estadístico de los datos antes de comercializar cualquier producto nuevo. Considere un estudio que incluye tres tipos de pastas (delgada, delgada con ajo y orégano, y delgada con trozos de queso). Dom's también está estudiando tres salsas (estándar, una nueva salsa con más ajo y una nueva salsa con albahaca fresca).

a) ¿Cuántas combinaciones de pasta y salsa se incluyen?
b) ¿Cuál es la probabilidad de que un juez reciba una pasta delgada sencilla con salsa estándar en su primera prueba de sabor?

2.63 A continuación se listan los porcentajes, proporcionados por *Consumer Digest* (julio/agosto de 1996), de las probables ubicaciones de las PC en una casa:

Dormitorio de adultos:	0.03
Dormitorio de niños:	0.15
Otro dormitorio:	0.14
Oficina o estudio:	0.40
Otra habitación:	0.28

a) ¿Cuál es la probabilidad de que una PC esté en un dormitorio?
b) ¿Cuál es la probabilidad de que no esté en un dormitorio?
c) Suponga que de entre las casas que tienen una PC se selecciona una al azar, ¿en qué habitación esperaría encontrar una PC?

2.64 Existe interés por la vida de un componente electrónico. Suponga que se sabe que la probabilidad de que el componente funcione más de 6000 horas es 0.42. Suponga, además, que la probabilidad de que el componente *no dure más de* 4000 horas es 0.04.

a) ¿Cuál es la probabilidad de que la vida del componente sea menor o igual a 6000 horas?
b) ¿Cuál es la probabilidad de que la vida del componente sea mayor que 4000 horas?

2.65 Considere la situación del ejercicio 2.64. Sea *A* el evento de que el componente falle en una prueba específica y *B* el evento de que se deforme pero no falle. El evento *A* ocurre con una probabilidad de 0.20 y el evento *B* ocurre con una probabilidad de 0.35.

a) ¿Cuál es la probabilidad de que el componente no falle en la prueba?
b) ¿Cuál es la probabilidad de que el componente funcione perfectamente bien (es decir, que ni se deforme ni falle en la prueba)?
c) ¿Cuál es la probabilidad de que el componente falle o se deforme en la prueba?

2.66 A los obreros de las fábricas se les motiva constantemente a practicar la tolerancia cero para prevenir accidentes en el lugar de trabajo. Los accidentes pueden ocurrir porque el ambiente o las condiciones laborales son inseguros. Por otro lado, los accidentes pueden ocurrir por negligencia o fallas humanas. Además, los horarios de trabajo de 7:00 A.M. a 3:00 P.M. (turno matutino), de 3:00 P.M. a 11:00 P.M. (turno vespertino) y de 11:00 P.M. a 7:00 A.M. (turno nocturno) podría ser un factor. El año pasado ocurrieron 300 accidentes. Los porcentajes de los accidentes por la combinación de condiciones son los que siguen:

Turno	Condiciones inseguras	Fallas humanas
Matutino	5%	32%
Vespertino	6%	25%
Nocturno	2%	30%

Si se elige aleatoriamente un reporte de accidente de entre los 300 reportes,
a) ¿Cuál es la probabilidad de que el accidente haya ocurrido en el turno nocturno?
b) ¿Cuál es la probabilidad de que el accidente haya ocurrido debido a una falla humana?
c) ¿Cuál es la probabilidad de que el accidente haya ocurrido debido a las condiciones inseguras?
d) ¿Cuál es la probabilidad de que el accidente haya ocurrido durante los turnos vespertino o nocturno?

2.67 Considere la situación del ejemplo 2.32 de la página 58.
a) ¿Cuál es la probabilidad de que el número de automóviles que recibirán servicio del mecánico no sea mayor de 4?
b) ¿Cuál es la probabilidad de que el mecánico dé servicio a menos de 8 automóviles?
c) ¿Cuál es la probabilidad de que el mecánico dé servicio a 3 o 4 automóviles?

2.68 Existe interés por el tipo de horno, eléctrico o de gas, que se compra en una tienda departamental específica. Considere la decisión que al respecto toman seis clientes distintos.
a) Suponga que hay 0.40 de probabilidades de que como máximo dos de esos clientes compren un horno eléctrico. ¿Cuál será la probabilidad de que al menos tres compren un horno eléctrico?
b) Suponga que se sabe que la probabilidad de que los seis compren el horno eléctrico es 0.007, mientras que la probabilidad de que los seis compren el horno de gas es 0.104. ¿Cuál es la probabilidad de vender, por lo menos, un horno de cada tipo?

2.69 En muchas áreas industriales es común que se utilicen máquinas para llenar las cajas de productos. Esto ocurre tanto en la industria de comestibles como en otras que fabrican productos de uso doméstico, como los detergentes. Dichas máquinas no son perfectas y, de hecho, podrían cumplir las especificaciones de llenado de las cajas (A), llenarlas por debajo del nivel especificado (B) o rebasar el límite de llenado (C). Por lo general, lo que se busca evitar es la práctica del llenado insuficiente. Sea $P(B) = 0.001$, mientras que $P(A) = 0.990$.
a) Determine $P(C)$.
b) ¿Cuál es la probabilidad de que la máquina no llene de manera suficiente?
c) ¿Cuál es la probabilidad de que la máquina llene de más o de menos?

2.70 Considere la situación del ejercicio 2.69. Suponga que se producen 50,000 cajas de detergente por semana, y que los clientes "devuelven" las cajas que no están suficientemente llenas y solicitan que se les reembolse lo que pagaron por ellas. Suponga que se sabe que el "costo" de producción de cada caja es de $4.00 y que se venden a $4.50.
a) ¿Cuál es la utilidad semanal cuando no hay devoluciones de cajas defectuosas?
b) ¿Cuál es la pérdida en utilidades esperada debido a la devolución de cajas insuficientemente llenadas?

2.71 Como podría sugerir la situación del ejercicio 2.69, a menudo los procedimientos estadísticos se utilizan para control de calidad (es decir, control de calidad industrial). A veces el *peso* de un producto es una variable importante que hay que controlar. Se dan especificaciones de peso para ciertos productos empacados, y si un paquete no las cumple (está muy ligero o muy pesado) se rechaza. Los datos históricos sugieren que la probabilidad de que un producto empacado cumpla con las especificaciones de peso es 0.95; mientras que la probabilidad de que sea demasiado ligero es 0.002. El fabricante invierte $20.00 en la producción de cada uno de los productos empacados y el consumidor los adquiere a un precio de $25.00.
a) ¿Cuál es la probabilidad de que un paquete elegido al azar de la línea de producción sea demasiado pesado?
b) Si todos los paquetes cumplen con las especificaciones de peso, ¿qué utilidad recibirá el fabricante por cada 10,000 paquetes que venda?

c) Suponga que todos los paquetes defectuosos fueron rechazados y perdieron todo su valor, ¿a cuánto se reduciría la utilidad de la venta de 10,000 paquetes debido a que no se cumplieron las especificaciones de peso?

2.72 Demuestre que

$$P(A' \cap B') = 1 + P(A \cap B) - P(A) - P(B).$$

2.6 Probabilidad condicional, independencia y regla del producto

Un concepto muy importante en la teoría de probabilidad es la probabilidad condicional. En algunas aplicaciones el profesional se interesa por la estructura de probabilidad bajo ciertas restricciones. Por ejemplo, en epidemiología, en lugar de estudiar las probabilidades de que una persona de la población general tenga diabetes, podría ser más interesante conocer esta probabilidad en un grupo distinto, como el de las mujeres asiáticas cuya edad está en el rango de 35 a 50 años, o como el de los hombres hispanos cuya edad está entre los 40 y los 60 años. A este tipo de probabilidad se le conoce como probabilidad condicional.

Probabilidad condicional

La probabilidad de que ocurra un evento B cuando se sabe que ya ocurrió algún evento A se llama **probabilidad condicional** y se denota con $P(B|A)$. El símbolo $P(B|\mathrm{A})$ por lo general se lee como "la probabilidad de que ocurra B, dado que ocurrió A", o simplemente, "la probabilidad de B, dado A".

Considere el evento B de obtener un cuadrado perfecto cuando se lanza un dado. El dado se construye de modo que los números pares tengan el doble de probabilidad de ocurrencia que los números nones. Con base en el espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$, en el que a los números impares y a los pares se les asignaron probabilidades de 1/9 y 2/9, respectivamente, la probabilidad de que ocurra B es de 1/3. Suponga ahora que se sabe que el lanzamiento del dado tiene como resultado un número mayor que 3. Tenemos ahora un espacio muestral reducido, $A = \{4, 5, 6\}$, que es un subconjunto de S. Para encontrar la probabilidad de que ocurra B, en relación con el espacio muestral A, debemos comenzar por asignar nuevas probabilidades a los elementos de A, que sean proporcionales a sus probabilidades originales de modo que su suma sea 1. Al asignar una probabilidad de w al número non en A y una probabilidad de $2w$ a los dos números pares, tenemos $5w = 1$ o $w = 1/5$. En relación con el espacio A, encontramos que B contiene sólo el elemento 4. Si denotamos este evento con el símbolo $B|A$, escribimos $B \mid A = \{4\}$ y, en consecuencia,

$$P(B \mid A) = \frac{2}{5}.$$

Este ejemplo ilustra que los eventos pueden tener probabilidades diferentes cuando se consideran en relación con diferentes espacios muestrales.

También podemos escribir

$$P(B \mid A) = \frac{2}{5} = \frac{2/9}{5/9} = \frac{P(A \cap B)}{P(A)},$$

donde $P(A \cap B)$ y $P(A)$ se calculan a partir del espacio muestral original S. En otras palabras, una probabilidad condicional relativa a un subespacio A de S se puede calcular en forma directa de las probabilidades que se asignan a los elementos del espacio muestral original S.

Definición 2.10: La probabilidad condicional de *B,* dado *A,* que se denota con $P(B|A)$, se define como

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

, siempre que $P(A) > 0$.

Un ejemplo más: suponga que tenemos un espacio muestral S constituido por la población de adultos de una pequeña ciudad que cumplen con los requisitos para obtener un título universitario. Debemos clasificarlos de acuerdo con su género y situación laboral. Los datos se presentan en la tabla 2.1.

Tabla 2.1: Clasificación de los adultos de una pequeña ciudad

	Empleado	**Desempleado**	**Total**
Hombre	460	40	500
Mujer	140	260	400
Total	600	300	900

Se seleccionará al azar a uno de estos individuos para que realice un viaje a través del país con el fin de promover las ventajas de establecer industrias nuevas en la ciudad. Nos interesaremos en los eventos siguientes:

M: se elige a un hombre,

E: el elegido tiene empleo.

Al utilizar el espacio muestral reducido *E,* encontramos que

$$P(M|E) = \frac{460}{600} = \frac{23}{30}.$$

Sea $n(A)$ el número de elementos en cualquier conjunto A. Podemos utilizar esta notación, puesto que cada uno de los adultos tiene las mismas probabilidades de ser elegido, para escribir

$$P(M|E) = \frac{n(E \cap M)}{n(E)} = \frac{n(E \cap M)/n(S)}{n(E)/n(S)} = \frac{P(E \cap M)}{P(E)},$$

en donde $P(E \cap M)$ y $P(E)$ se calculan a partir del espacio muestral original S. Para verificar este resultado observe que

$$P(E) = \frac{600}{900} = \frac{2}{3} \quad \text{y} \quad P(E \cap M) = \frac{460}{900} = \frac{23}{45}.$$

Por lo tanto,

$$P(M|E) = \frac{23/45}{2/3} = \frac{23}{30},$$

como antes.

Ejemplo 2.34: La probabilidad de que un vuelo programado normalmente salga a tiempo es $P(D) = 0.83$, la probabilidad de que llegue a tiempo es $P(A) = 0.82$ y la probabilidad de que

salga y llegue a tiempo es $P(D \cap A) = 0.78$. Calcule la probabilidad de que un avión *a*) llegue a tiempo, dado que salió a tiempo; y *b*) salió a tiempo, dado que llegó a tiempo.

***Solución*:** Al utilizar la definición 2.10 tenemos lo que sigue:

a) La probabilidad de que un avión llegue a tiempo, dado que salió a tiempo es

$$P(A|D) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.94.$$

b) La probabilidad de que un avión haya salido a tiempo, dado que llegó a tiempo es

$$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.95.$$

La noción de probabilidad condicional brinda la capacidad de reevaluar la idea de probabilidad de un evento a la luz de la información adicional; es decir, cuando se sabe que ocurrió otro evento. La probabilidad $P(A|B)$ es una actualización de $P(A)$ basada en el conocimiento de que ocurrió el evento B. En el ejemplo 2.34 es importante conocer la probabilidad de que el vuelo llegue a tiempo. Tenemos la información de que el vuelo no salió a tiempo. Con esta información adicional, la probabilidad más pertinente es $P(A|D')$, esto es, la probabilidad de que llegue a tiempo, dado que no salió a tiempo. A menudo las conclusiones que se obtienen a partir de observar la probabilidad condicional más importante cambian drásticamente la situación. En este ejemplo, el cálculo de $P(A|D')$ es

$$P(A|D') = \frac{P(A \cap D')}{P(D')} = \frac{0.82 - 0.78}{0.17} = 0.24.$$

Como resultado, la probabilidad de una llegada a tiempo disminuye significativamente ante la presencia de la información adicional.

Ejemplo 2.35: El concepto de probabilidad condicional tiene innumerables aplicaciones industriales y biomédicas. Considere un proceso industrial en el ramo textil, en el que se producen listones de una tela específica. Los listones pueden resultar con defectos en dos de sus características: la longitud y la textura. En el segundo caso el proceso de identificación es muy complicado. A partir de información histórica del proceso se sabe que 10% de los listones no pasan la prueba de longitud, que 5% no pasan la prueba de textura y que sólo 0.8% no pasan ninguna de las dos pruebas. Si en el proceso se elige un listón al azar y una medición rápida identifica que no pasa la prueba de longitud, ¿cuál es la probabilidad de que la textura esté defectuosa?

***Solución*:** Considere los eventos

L: defecto en longitud, T: defecto en textura.

Dado que el listón tiene una longitud defectuosa, la probabilidad de que este listón tenga una textura defectuosa está dada por

$$P(T|L) = \frac{P(T \cap L)}{P(L)} = \frac{0.008}{0.1} = 0.08.$$

Eventos independientes

En el experimento del lanzamiento de un dado de la página 62 señalamos que $P(B|A) = 2/5$, mientras que $P(B) = 1/3$. Es decir, $P(B|A) \neq P(B)$, lo cual indica que B depende de A. Consideremos ahora un experimento en el que se sacan 2 cartas, una después de la otra, de una baraja ordinaria, con reemplazo. Los eventos se definen como

A: la primera carta es un as,

B: la segunda carta es una espada.

Como la primera carta se reemplaza, nuestro espacio muestral para la primera y segunda cartas consta de 52 cartas, que contienen 4 ases y 13 espadas. Entonces,

$$P(B|A) = \frac{13}{52} = \frac{1}{4} \quad \text{y} \quad P(B) = \frac{13}{52} = \frac{1}{4}.$$

Es decir, $P(B|A) = P(B)$. Cuando esto es cierto, se dice que los eventos A y B son **independientes**.

Aunque la probabilidad condicional permite alterar la probabilidad de un evento a la luz de material adicional, también nos permite entender mejor el muy importante concepto de **independencia** o, en el contexto actual, de eventos independientes. En el ejemplo 2.34 del aeropuerto, $P(A|D)$ difiere de $P(A)$. Esto sugiere que la ocurrencia de D influye en A y esto es lo que, de hecho, se espera en este caso. Sin embargo, considere la situación en donde tenemos los eventos A y B, y

$$P(A|B) = P(A).$$

En otras palabras, la ocurrencia de B no influye en las probabilidades de ocurrencia de A. Aquí la ocurrencia de A es independiente de la ocurrencia de B. No podemos dejar de resaltar la importancia del concepto de independencia, ya que desempeña un papel vital en el material de casi todos los capítulos de este libro y en todas las áreas de la estadística aplicada.

Definición 2.11: Dos eventos A y B son **independientes** si y sólo si

$$P(B|A) = P(B) \text{ o } P(A|B) = P(A),$$

si se asume la existencia de probabilidad condicional. De otra forma, A y B son **dependientes**.

La condición $P(B|A) = P(B)$ implica que $P(A|B) = P(A)$, y viceversa. Para los experimentos de extracción de una carta, donde mostramos que $P(B|A) = P(B) = 1/4$, también podemos ver que $P(A|B) = P(A) = 1/13$.

La regla de producto o regla multiplicativa

Al multiplicar la fórmula de la definición 2.10 por $P(A)$, obtenemos la siguiente **regla multiplicativa** importante (o **regla de producto**), que nos permite calcular la probabilidad de que ocurran dos eventos.

Teorema 2.10: Si en un experimento pueden ocurrir los eventos A y B, entonces

$$P(A \cap B) = P(A)P(B|A), \text{ siempre que } P(A) > 0.$$

Por consiguiente, la probabilidad de que ocurran A y B es igual a la probabilidad de que ocurra A multiplicada por la probabilidad condicional de que ocurra B, dado que ocurre A. Como los eventos $A \cap B$ y $B \cap A$ son equivalentes, del teorema 2.10 se deduce que también podemos escribir

$$P(A \cap B) = P(B \cap A) = P(B)P(A|B).$$

En otras palabras, no importa qué evento se considere como A ni qué evento se considere como B.

Ejemplo 2.36: Suponga que tenemos una caja de fusibles que contiene 20 unidades, de las cuales 5 están defectuosas. Si se seleccionan 2 fusibles al azar y se retiran de la caja, uno después del otro, sin reemplazar el primero, ¿cuál es la probabilidad de que ambos fusibles estén defectuosos?

***Solución*:** Sean A el evento de que el primer fusible esté defectuoso y B el evento de que el segundo esté defectuoso; entonces, interpretamos $A \cap B$ como el evento de que ocurra A, y entonces B ocurre después de que haya ocurrido A. La probabilidad de sacar primero un fusible defectuoso es 1/4; entonces, la probabilidad de separar un segundo fusible defectuoso de los restantes 4 es 4/19. Por lo tanto,

$$P(A \cap B) = \left(\frac{1}{4}\right)\left(\frac{4}{19}\right) = \frac{1}{19}.$$

Ejemplo 2.37: Una bolsa contiene 4 bolas blancas y 3 negras, y una segunda bolsa contiene 3 blancas y 5 negras. Se saca una bola de la primera bolsa y se coloca sin verla en la segunda bolsa. ¿Cuál es la probabilidad de que ahora se saque una bola negra de la segunda bolsa?

***Solución*:** N_1, N_2 y B_1 representan, respectivamente, la extracción de una bola negra de la bolsa 1, una bola negra de la bolsa 2 y una bola blanca de la bolsa 1. Nos interesa la unión de los eventos mutuamente excluyentes $N_1 \cap N_2$ y $B_1 \cap N_2$. Las diversas posibilidades y sus probabilidades se ilustran en la figura 2.8. Entonces

$$\begin{aligned} P[(N_1 \cap N_2) \text{ o } (B_1 \cap N_2)] &= P(N_1 \cap N_2) + P(B_1 \cap N_2) \\ &= P(N_1)P(N_2|N_1) + P(B_1)P(N_2|B_1) \\ &= \left(\frac{3}{7}\right)\left(\frac{6}{9}\right) + \left(\frac{4}{7}\right)\left(\frac{5}{9}\right) = \frac{38}{63}. \end{aligned}$$

Si, en el ejemplo 2.36, el primer fusible se reemplaza y los fusibles se reacomodan por completo antes de extraer el segundo, entonces la probabilidad de que se extraiga un fusible defectuoso en la segunda selección sigue siendo 1/4; es decir, $P(B|A) = P(B)$, y los eventos A y B son independientes. Cuando esto es cierto podemos sustituir $P(B)$ por $P(B|A)$ en el teorema 2.10 para obtener la siguiente regla multiplicativa especial.

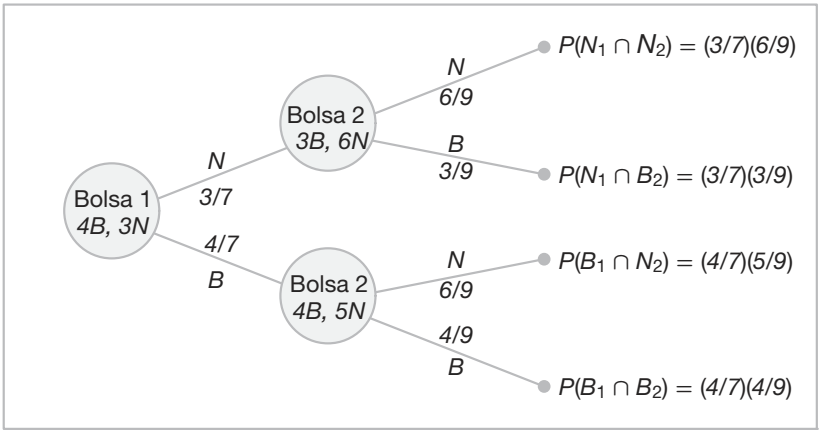


Figura 2.8: Diagrama de árbol para el ejemplo 2.37.

Teorema 2.11: Dos eventos A y B son independientes si y sólo si

$$P(A \cap B) = P(A)P(B).$$

Por lo tanto, para obtener la probabilidad de que ocurran dos eventos independientes simplemente calculamos el producto de sus probabilidades individuales.

Ejemplo 2.38: Una pequeña ciudad dispone de un carro de bomberos y una ambulancia para emergencias. La probabilidad de que el carro de bomberos esté disponible cuando se necesite es 0.98 y la probabilidad de que la ambulancia esté disponible cuando se le requiera es 0.92. En el evento de un herido en un incendio, calcule la probabilidad de que tanto la ambulancia como el carro de bomberos estén disponibles, suponiendo que operan de forma independiente.

***Solución*:** Sean A y B los respectivos eventos de que estén disponibles el carro de bomberos y la ambulancia. Entonces,

$$P(A \cap B) = P(A)P(B) = (0.98)(0.92) = 0.9016.$$

Ejemplo 2.39: Un sistema eléctrico consta de cuatro componentes, como se ilustra en la figura 2.9. El sistema funciona si los componentes A y B funcionan, y si funciona cualquiera de los componentes C o D. La confiabilidad (probabilidad de que funcionen) de cada uno de los componentes también se muestra en la figura 2.9. Calcule la probabilidad de *a*) que el sistema completo funcione y de *b*) que el componente C no funcione, dado que el sistema completo funciona. Suponga que los cuatro componentes funcionan de manera independiente.

***Solución*:** En esta configuración del sistema, A, B y el subsistema C y D constituyen un sistema de circuitos en serie; mientras que el subsistema C y D es un sistema de circuitos en paralelo.

a) Es evidente que la probabilidad de que el sistema completo funcione se puede calcular de la siguiente manera:

$$\begin{aligned} P[A \cap B \cap (C \cup D)] &= P(A)P(B)P(C \cup D) = P(A)P(B)[1 - P(C' \cap D')] \\ &= P(A)P(B)[1 - P(C')P(D')] \\ &= (0.9)(0.9)[1 - (1 - 0.8)(1 - 0.8)] = 0.7776. \end{aligned}$$

Las igualdades anteriores son válidas debido a la independencia entre los cuatro componentes.

b) Para calcular la probabilidad condicional en este caso, observe que

$$\begin{aligned} P &= \frac{P\,(\text{el sistema funciona pero } C \text{ no funciona})}{P\,(\text{el sistema funciona})} \\ &= \frac{P\,(A \cap B \cap C' \cap D)}{P\,(\text{el sistema funciona})} = \frac{(0.9)(0.9)(1 - 0.8)(0.8)}{0.7776} = 0.1667. \end{aligned}$$

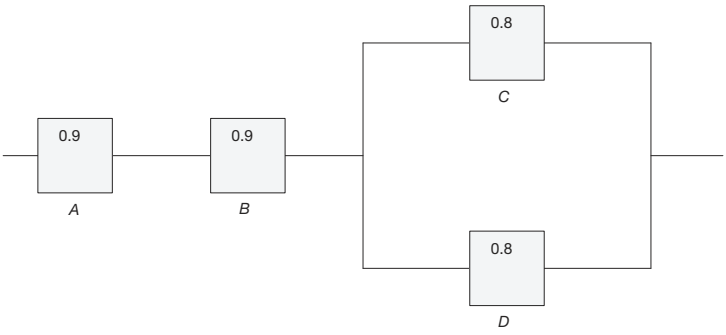


Figura 2.9: Un sistema eléctrico para el ejemplo 2.39.

La regla multiplicativa se puede extender a situaciones con más de dos eventos.

Teorema 2.12: Si, en un experimento, pueden ocurrir los eventos $A_1, A_2, \ldots, A_k$, entonces

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_k) \\ &\qquad = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdots P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1}). \end{aligned}$$

Si los eventos $A_1, A_2, \ldots, A_k$ son independientes, entonces

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)P(A_2)\cdots P(A_k)$$

Ejemplo 2.40: Se sacan tres cartas seguidas, sin reemplazo, de una baraja ordinaria. Encuentre la probabilidad de que ocurra el evento $A_1 \cap A_2 \cap A_3$, donde A_1 es el evento de que la primera carta sea un as rojo, A_2 el evento de que la segunda carta sea un 10 o una jota y A_3 el evento de que la tercera carta sea mayor que 3 pero menor que 7.

***Solución*:** Primero definimos los eventos:

A_1: la primera carta es un as rojo,

A_2: la segunda carta es un 10 o una jota,

A_3: la tercera carta es mayor que 3 pero menor que 7.

Ahora bien,

$$P(A_1) = \frac{2}{52},\ P(A_2|A_1) = \frac{8}{51},\ P(A_3|A_1 \cap A_2) = \frac{12}{50},$$

por lo tanto, por medio del teorema 2.12,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \\ &= \left(\frac{2}{52}\right)\left(\frac{8}{51}\right)\left(\frac{12}{50}\right) = \frac{8}{5525}. \end{aligned}$$

La propiedad de independencia establecida en el teorema 2.11 se puede extender a situaciones con más de dos eventos. Considere, por ejemplo, el caso de los tres eventos A, B y C. No basta con tener $P(A \cap B \cap C) = P(A)P(B)P(C)$ como una definición de independencia entre los tres. Suponga que $A = B$ y $C = \emptyset$, el conjunto vacío. Aunque $A \cap B \cap C = f$, que da como resultado $P(A \cap B \cap C) = 0 = P(A)P(B)P(C)$, los eventos A y B no son independientes. En consecuencia, tenemos la siguiente definición:

Definición 2.12: Un conjunto de eventos $\mathcal{A} = \{A_1, \ldots, A_n\}$ son mutuamente independientes si para cualquier subconjunto de $\mathcal{A}$, A_{i1} ..., A_{ik}, para $k \leq n$, tenemos

$$P(A_{i1} \cap \cdots \cap A_{ik}) = P(A_{i1}) \cdots P(A_{ik}).$$

Ejercicios

2.73 Si R es el evento de que un convicto cometa un robo a mano armada y D es el evento de que el convicto venda drogas, exprese en palabras lo que en probabilidades se indica como

a) $P(R|D)$;
b) $P(D'|R)$;
c) $P(R'|D')$.

2.74 Un grupo de estudiantes de física avanzada se compone de 10 alumnos de primer año, 30 del último año y 10 graduados. Las calificaciones finales muestran que 3 estudiantes de primer año, 10 del último año y 5 de los graduados obtuvieron 10 en el curso. Si se elige un estudiante al azar de este grupo y se descubre que es uno de los que obtuvieron 10 de calificación, ¿cuál es la probabilidad de que sea un estudiante de último año?

2.75 La siguiente es una clasificación, según el género y el nivel de escolaridad, de una muestra aleatoria de 200 adultos.

Escolaridad	Hombre	Mujer
Primaria	38	45
Secundaria	28	50
Universidad	22	17

Si se elige una persona al azar de este grupo, ¿cuál es la probabilidad de que…

a) la persona sea hombre, dado que su escolaridad es de secundaria?;
b) la persona no tenga un grado universitario, dado que es mujer?

2.76 En un experimento para estudiar la relación que existe entre el hábito de fumar y la hipertensión arterial se reúnen los siguientes datos para 180 individuos:

	No fumadores	Fumadores moderados	Fumadores empedernidos
H	21	36	30
SH	48	26	19

donde las letras *H* y *SH* de la tabla representan *Hipertensión* y *Sin hipertensión*, respectivamente. Si se selecciona uno de estos individuos al azar, calcule la probabilidad de que la persona…

a) sufra hipertensión, dado que es una fumadora empedernida;
b) no fume, dado que no padece hipertensión.

2.77 En un grupo de 100 estudiantes de bachillerato que están cursando el último año, 42 cursaron matemáticas, 68 psicología, 54 historia, 22 matemáticas e historia, 25 matemáticas y psicología, 7 historia pero ni matemáticas ni psicología, 10 las tres materias y 8 no cursaron ninguna de las tres. Seleccione al azar a un

estudiante de este grupo y calcule la probabilidad de los siguientes eventos:

a) Una persona inscrita en psicología y cursa las tres materias;
b) Una persona que no está inscrita en psicología y esté cursando historia y matemáticas.

2.78 Un fabricante de una vacuna para la gripe está interesado en determinar la calidad de su suero. Con ese fin tres departamentos diferentes procesan los lotes de suero y tienen tasas de rechazo de 0.10, 0.08 y 0.12, respectivamente. Las inspecciones de los tres departamentos son secuenciales e independientes.

a) ¿Cuál es la probabilidad de que un lote de suero sobreviva a la primera inspección departamental pero sea rechazado por el segundo departamento?
b) ¿Cuál es la probabilidad de que un lote de suero sea rechazado por el tercer departamento?

2.79 En *USA Today* (5 de septiembre de 1996) se listaron los siguientes resultados de una encuesta sobre el uso de ropa para dormir mientras se viaja:

	Hombre	Mujer	Total
Ropa interior	0.020	0.024	0.244
Camisón	0.002	0.180	0.182
Nada	0.160	0.018	0.178
Pijama	0.102	0.073	0.175
Camiseta	0.046	0.088	0.134
Otros	0.084	0.003	0.087

a) ¿Cuál es la probabilidad de que un viajero sea una mujer que duerme desnuda?
b) ¿Cuál es la probabilidad de que un viajero sea hombre?
c) Si el viajero fuera hombre, ¿cuál sería la probabilidad de que duerma con pijama?
d) ¿Cuál es la probabilidad de que un viajero sea hombre si duerme con pijama o con camiseta?

2.80 La probabilidad de que cuando se tenga que llenar el tanque de gasolina de un automóvil también se necesite cambiarle el aceite es 0.25, la probabilidad de que también se le tenga que cambiar el filtro de aceite es 0.40, y la probabilidad de que se necesite cambiarle el aceite y el filtro es 0.14.

a) Si se le tiene que cambiar el aceite, ¿cuál es la probabilidad de que también se necesite cambiarle el filtro?
b) Si se le tiene que cambiar el filtro de aceite, ¿cuál es la probabilidad de que también se le tenga que cambiar el aceite?

2.81 La probabilidad de que un hombre casado vea cierto programa de televisión es 0.4 y la probabilidad de que lo vea una mujer casada es 0.5. La probabilidad de que un hombre vea el programa, dado que su esposa lo ve, es 0.7. Calcule la probabilidad de que

a) una pareja casada vea el programa;
b) una esposa vea el programa dado que su esposo lo ve;
c) al menos uno de los miembros de la pareja casada vea el programa.

2.82 Para parejas casadas que viven en cierto suburbio, la probabilidad de que el esposo vote en un referéndum es 0.21, la probabilidad de que vote la esposa es 0.28 y la probabilidad de que ambos voten es 0.15. ¿Cuál es la probabilidad de que…

a) al menos uno de los miembros de la pareja casada vote?
b) una esposa vote, dado que su esposo vota?
c) un esposo vote, dado que su esposa no vota?

2.83 La probabilidad de que un vehículo que entra a las Cavernas Luray tenga matrícula de Canadá es 0.12, la probabilidad de que sea una casa rodante es 0.28 y la probabilidad de que sea una casa rodante con matrícula de Canadá es 0.09. ¿Cuál es la probabilidad de que…

a) una casa rodante que entra a las Cavernas Luray tenga matrícula de Canadá?
b) un vehículo con matrícula de Canadá que entra a las Cavernas Luray sea una casa rodante?
c) un vehículo que entra a las Cavernas Luray no tenga matrícula de Canadá o no sea una casa rodante?

2.84 La probabilidad de que el jefe de familia esté en casa cuando llame el representante de marketing de una empresa es 0.4. Dado que el jefe de familia está en casa, la probabilidad de que la empresa le venda un producto es 0.3. Encuentre la probabilidad de que el jefe de familia esté en casa y compre productos de la empresa.

2.85 La probabilidad de que un doctor diagnostique de manera correcta una enfermedad específica es 0.7. Dado que el doctor hace un diagnóstico incorrecto, la probabilidad de que el paciente entable una demanda legal es 0.9. ¿Cuál es la probabilidad de que el doctor haga un diagnóstico incorrecto y el paciente lo demande?

2.86 En 1970, 11% de los estadounidenses completaron cuatro años de universidad; de ese porcentaje 43 % eran mujeres. En 1990, 22% de los estadounidenses completaron cuatro años de universidad, un porcentaje del cual 53 % fueron mujeres. (*Time,* 19 de enero de 1996).

a) Dado que una persona completó cuatro años de universidad en 1970, ¿cuál es la probabilidad de que esa persona sea mujer?

b) ¿Cuál es la probabilidad de que una mujer haya terminado cuatro años de universidad en 1990?
c) ¿Cuál es la probabilidad de que en 1990 un hombre no haya terminado la universidad?

2.87 Un agente de bienes raíces tiene 8 llaves maestras para abrir varias casas nuevas. Sólo 1 llave maestra abrirá cualquiera de las casas. Si 40% de estas casas por lo general se dejan abiertas, ¿cuál es la probabilidad de que el agente de bienes raíces pueda entrar en una casa específica, si selecciona 3 llaves maestras al azar antes de salir de la oficina?

2.88 Antes de la distribución de cierto software estadístico se prueba la precisión de cada cuarto disco compacto (CD). El proceso de prueba consiste en correr cuatro programas independientes y verificar los resultados. La tasa de falla para los 4 programas de prueba son 0.01, 0.03, 0.02 y 0.01, respectivamente.
a) ¿Cuál es la probabilidad de que uno de los CD que se pruebe no pase la prueba?
b) Dado que se prueba un CD, ¿cuál es la probabilidad de que falle el programa 2 o 3?
c) En una muestra de 100, ¿cuántos CD esperaría que se rechazaran?
d) Dado que un CD está defectuoso, ¿cuál es la probabilidad de que se pruebe?

2.89 Una ciudad tiene dos carros de bomberos que operan de forma independiente. La probabilidad de que un carro específico esté disponible cuando se le necesite es 0.96.
a) ¿Cuál es la probabilidad de que ninguno esté disponible cuando se necesite?
b) ¿Cuál es la probabilidad de que un carro de bomberos esté disponible cuando se le necesite?

2.90 La contaminación de los ríos de Estados Unidos ha sido un problema por muchos años. Considere los siguientes eventos:

A: el río está contaminado.
B: al probar una muestra de agua se detecta contaminación.
C: se permite pescar.

Suponga que $P(A) = 0.3$, $P(B|A) = 0.75$, $P(B|A') = 0.20$, $P(C|A \cap B) = 0.20$, $P(C|A' \cap B) = 0.15$, $P(C|A \cap B') = 0.80$ y $P(C|A' \cap B') = 0.90$.
a) Calcule $P(A \cap B \cap C)$.
b) Calcule $P(B' \cap C)$.
c) Calcule $P(C)$.
d) Calcule la probabilidad de que el río esté contaminado, dado que está permitido pescar y que la muestra probada no detectó contaminación.

2.91 Encuentre la posibilidad de seleccionar aleatoriamente 4 litros de leche en buenas condiciones sucesivamente de un refrigerador que contiene 20 litros, de los cuales 5 están echados a perder, utilizando
a) la primera fórmula del teorema 2.12 de la página 68;
b) las fórmulas del teorema 2.6 y la regla 2.3 de las páginas 50 y 54, respectivamente.

2.92 Imagine el diagrama de un sistema eléctrico como el que se muestra en la figura 2.10. ¿Cuál es la probabilidad de que el sistema funcione? Suponga que los componentes fallan de forma independiente.

2.93 En la figura 2.11 se muestra un sistema de circuitos. Suponga que los componentes fallan de manera independiente.
a) ¿Cuál es la probabilidad de que el sistema completo funcione?
b) Dado que el sistema funciona, ¿cuál es la probabilidad de que el componente A no funcione?

2.94 En la situación del ejercicio 2.93 se sabe que el sistema no funciona. ¿Cuál es la probabilidad de que el componente A tampoco funcione?

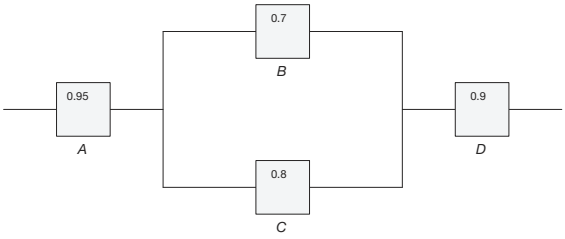


Figura 2.10: Diagrama para el ejercicio 2.92.

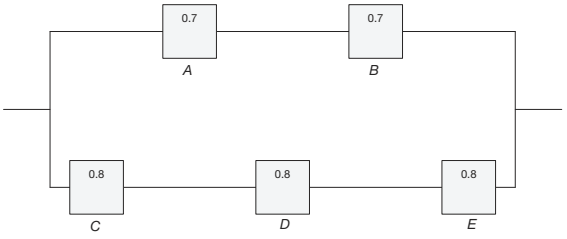


Figura 2.11: Diagrama para el ejercicio 2.93.

2.7 Regla de Bayes

La estadística bayesiana es un conjunto de herramientas que se utiliza en un tipo especial de inferencia estadística que se aplica en el análisis de datos experimentales en muchas situaciones prácticas de ciencia e ingeniería. La regla de Bayes es una de las normas más importantes de la teoría de probabilidad, ya que es el fundamento de la inferencia bayesiana, la cual se analizará en el capítulo 18.

Probabilidad total

Regresemos al ejemplo de la sección 2.6, en el que se selecciona un individuo al azar de entre los adultos de una pequeña ciudad para que viaje por el país promoviendo las ventajas de establecer industrias nuevas en la ciudad. Suponga que ahora se nos da la información adicional de que 36 de los empleados y 12 de los desempleados son miembros del Club Rotario. Deseamos encontrar la probabilidad del evento A de que el individuo seleccionado sea miembro del Club Rotario. Podemos remitirnos a la figura 2.12 y escribir A como la unión de los dos eventos mutuamente excluyentes $E \cap A$ y $E' \cap A$. Por lo tanto, $A = (E \cap A) \cup (E' \cap A)$, y mediante el corolario 2.1 del teorema 2.7 y luego mediante el teorema 2.10, podemos escribir

$$\begin{aligned} P(A) &= P\left[(E \cap A) \cup (E' \cap A)\right] = P(E \cap A) + P(E' \cap A) \\ &= P(E)P(A|E) + P(E')P(A|E'). \end{aligned}$$

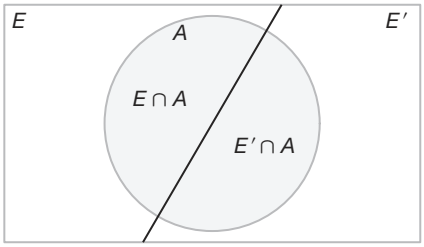


Figura 2.12: Diagrama de Venn para los eventos A, E y E'.

Los datos de la sección 2.6, junto con los datos adicionales antes dados para el conjunto A, nos permiten calcular

$$P(E) = \frac{600}{900} = \frac{2}{3}, \quad P(A|E) = \frac{36}{600} = \frac{3}{50},$$

y

$$P(E') = \frac{1}{3}, \quad P(A|E') = \frac{12}{300} = \frac{1}{25}.$$

Si mostramos estas probabilidades mediante el diagrama de árbol de la figura 2.13, donde la primera rama da la probabilidad $P(E)P(A|E)$ y la segunda rama da la probabilidad

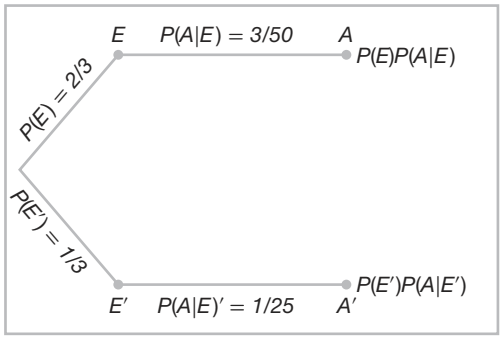


Figura 2.13: Diagrama de árbol para los datos de la página 63 con información adicional de la página 72.

la probabilidad $P(E')P(A|E')$, deducimos que

$$P(A) = \left(\frac{2}{3}\right)\left(\frac{3}{50}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{25}\right) = \frac{4}{75}.$$

Una generalización del ejemplo anterior para el caso en donde el espacio muestral se parte en k subconjuntos se cubre mediante el siguiente teorema, que algunas veces se denomina **teorema de probabilidad total** o **regla de eliminación**.

Teorema 2.13: Si los eventos $B_1, B_2, \ldots B_k$ constituyen una partición del espacio muestral S, tal que $P(B_i) \neq 0$ para $i = 1, 2, \ldots, k$, entonces, para cualquier evento A de S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

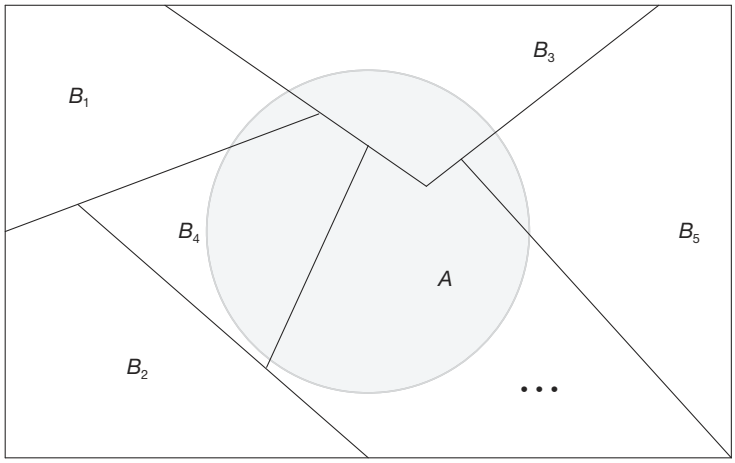


Figura 2.14: Partición del espacio muestral s.

***Prueba*:** Considere el diagrama de Venn de la figura 2.14. Se observa que el evento A es la unión de los eventos mutuamente excluyentes

$$B_1 \cap A,\ B_2 \cap A,\ldots,\ B_k \cap A;$$

es decir,

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A)$$

Por medio del corolario 2.2 del teorema 2.7 y el teorema 2.10 obtenemos

$$\begin{aligned} P(A) &= P[(B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A)] \\ &= P(B_1 \cap A) + P(B_2 \cap A) + \cdots + P(B_k \cap A) \\ &= \sum_{i=1}^{k} P(B_i \cap A) \\ &= \sum_{i=1}^{k} P(B_i)P(A|B_i). \end{aligned}$$

Ejemplo 2.41: Tres máquinas de cierta planta de ensamble, B_1, B_2 y B_3, montan 30%, 45% y 25% de los productos, respectivamente. Se sabe por experiencia que 2%, 3% y 2% de los productos ensamblados por cada máquina, respectivamente, tienen defectos. Ahora bien, suponga que se selecciona de forma aleatoria un producto terminado. ¿Cuál es la probabilidad de que esté defectuoso?

***Solución*:** Considere los siguientes eventos:

A: el producto está defectuoso,

B_1: el producto fue ensamblado con la máquina B_1,

B_2: el producto fue ensamblado con la máquina B_2,

B_3: el producto fue ensamblado con la máquina B_3.

Podemos aplicar la regla de eliminación y escribir

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3).$$

Si nos remitimos al diagrama de árbol de la figura 2.15 encontramos que las tres ramas dan las probabilidades

$$\begin{aligned} P(B_1)P(A|B_1) &= (0.3)(0.02) = 0.006, \\ P(B_2)P(A|B_2) &= (0.45)(0.03) = 0.0135, \\ P(B_3)P(A|B_3) &= (0.25)(0.02) = 0.005, \end{aligned}$$

en consecuencia,

$$P(A) = 0.006 + 0.0135 + 0.005 = 0.0245.$$

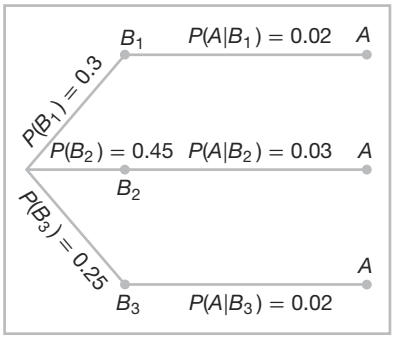


Figura 2.15: Diagrama de árbol para el ejemplo 2.41.

Regla de Bayes

Suponga que en lugar de calcular $P(A)$ mediante la regla de eliminación en el ejemplo 2.41, consideramos el problema de obtener la probabilidad condicional $P(B_i|A)$. En otras palabras, suponga que se selecciona un producto de forma aleatoria y que éste resulta defectuoso. ¿Cuál es la probabilidad de que este producto haya sido ensamblado con la máquina B_i? Las preguntas de este tipo se pueden contestar usando el siguiente teorema, denominado **regla de Bayes**:

Teorema 2.14: (**Regla de Bayes**) Si los eventos B_1, B_2,..., B_k constituyen una partición del espacio muestral S, donde $P(B_i) \neq 0$ para $i = 1, 2,...,k$, entonces, para cualquier evento A en S, tal que $P(A) \neq 0$,

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \quad \text{para} \quad r = 1, 2, \ldots, k.$$

***Prueba*:** Mediante la definición de probabilidad condicional,

$$P(B_r|A) = \frac{P(B_r \cap A)}{P(A)},$$

y después usando el teorema 2.13 en el denominador, tenemos

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)},$$

que completa la demostración. ∎

Ejemplo 2.42: Con referencia al ejemplo 2.41, si se elige al azar un producto y se encuentra que está defectuoso, ¿cuál es la probabilidad de que haya sido ensamblado con la máquina B_3?

***Solución*:** Podemos utilizar la regla de Bayes para escribir

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)},$$

y después al sustituir las probabilidades calculadas en el ejemplo 2.41, tenemos

$$P(B_3|A) = \frac{0.005}{0.006 + 0.0135 + 0.005} = \frac{0.005}{0.0245} = \frac{10}{49}.$$

En vista del hecho de que se seleccionó un producto defectuoso, este resultado sugiere que probablemente no fue ensamblado con la máquina B_3. ∎

Ejemplo 2.43: Una empresa de manufactura emplea tres planos analíticos para el diseño y desarrollo de un producto específico. Por razones de costos los tres se utilizan en momentos diferentes. De hecho, los planos 1, 2 y 3 se utilizan para 30%, 20% y 50% de los productos, respectivamente. La tasa de defectos difiere en los tres procedimientos de la siguiente manera,

$$P(D|P_1) = 0.01, \quad P(D|P_2) = 0.03, \quad P(D|P_3) = 0.02,$$

en donde $P(D|P_j)$ es la probabilidad de que un producto esté defectuoso, dado el plano j. Si se observa un producto al azar y se descubre que está defectuoso, ¿cuál de los planos tiene más probabilidades de haberse utilizado y, por lo tanto, de ser el responsable?

***Solución*:** A partir del planteamiento del problema

$$P(P_1) = 0.30, \quad P(P_2) = 0.20 \quad \text{y} \quad P(P_3) = 0.50,$$

debemos calcular $P(P_j|D)$ para $j = 1, 2, 3$. La regla de Bayes (teorema 2.14) muestra

$$P(P_1|D) = \frac{P(P_1)P(D|P_1)}{P(P_1)P(D|P_1) + P(P_2)P(D|P_2) + P(P_3)P(D|P_3)}$$
$$= \frac{(0.30)(0.01)}{(0.3)(0.01) + (0.20)(0.03) + (0.50)(0.02)} = \frac{0.003}{0.019} = 0.158.$$

De igual manera,

$$P(P_2|D) = \frac{(0.03)(0.20)}{0.019} = 0.316 \text{ y } P(P_3|D) = \frac{(0.02)(0.50)}{0.019} = 0.526.$$

La probabilidad condicional de un defecto, dado el plano 3, es la mayor de las tres; por consiguiente, un defecto en un producto elegido al azar tiene más probabilidad de ser el resultado de haber usado el plano 3. ∎

La regla de Bayes, un método estadístico llamado método bayesiano, ha adquirido muchas aplicaciones. En el capítulo 18 estudiaremos una introducción al método bayesiano.

Ejercicios

2.95 En cierta región del país se sabe por experiencia que la probabilidad de seleccionar un adulto mayor de 40 años de edad con cáncer es 0.05. Si la probabilidad de que un doctor diagnostique de forma correcta que una persona con cáncer tiene la enfermedad es 0.78, y la probabilidad de que diagnostique de forma incorrecta que una persona sin cáncer tiene la enfermedad es 0.06, ¿cuál es la probabilidad de que a un adulto mayor de 40 años se le diagnostique cáncer?

2.96 La policía planea hacer respetar los límites de velocidad usando un sistema de radar en 4 diferentes puntos a las orillas de la ciudad. Las trampas de radar en cada uno de los sitios L_1, L_2, L_3 y L_4 operarán 40%, 30 %, 20% y 30% del tiempo. Si una persona que excede el límite de velocidad cuando va a su trabajo tiene probabilidades de 0.2, 0.1, 0.5 y 0.2, respectivamente, de pasar por esos lugares, ¿cuál es la probabilidad de que reciba una multa por conducir con exceso de velocidad?

2.97 Remítase al ejercicio 2.95. ¿Cuál es la probabilidad de que una persona a la que se le diagnostica cáncer realmente tenga la enfermedad?

2.98 Si en el ejercicio 2.96 la persona es multada por conducir con exceso de velocidad en su camino al trabajo, ¿cuál es la probabilidad de que pase por el sistema de radar que se ubica en L_2?

2.99 Suponga que los cuatro inspectores de una fábrica de película colocan la fecha de caducidad en cada paquete de película al final de la línea de montaje. John, quien coloca la fecha de caducidad en 20% de los paquetes, no logra ponerla en uno de cada 200 paquetes; Tom, quien la coloca en 60% de los paquetes, no logra ponerla en uno de cada 100 paquetes; Jeff, quien la coloca en 15% de los paquetes, no lo hace una vez en cada 90 paquetes; y Pat, que fecha 5% de los paquetes, falla en uno de cada 200 paquetes. Si un cliente se queja de que su paquete de película no muestra la fecha de caducidad, ¿cuál es la probabilidad de que haya sido inspeccionado por John?

2.100 Una empresa telefónica regional opera tres estaciones de retransmisión idénticas en diferentes sitios. A continuación se muestra el número de desperfectos en cada estación reportados durante un año y las causas de éstos.

Estación	*A*	*B*	*C*
Problemas con el suministro de electricidad	2	1	1
Falla de la computadora	4	3	2
Fallas del equipo eléctrico	5	4	2
Fallas ocasionadas por otros errores humanos	7	5	5

Suponga que se reporta una falla y que se descubre que fue ocasionada por otros errores humanos. ¿Cuál es la probabilidad de que provenga de la estación *C*?

2.101 Una cadena de tiendas de pintura produce y vende pintura de látex y semiesmaltada. De acuerdo con las ventas a largo plazo, la probabilidad de que un cliente compre pintura de látex es 0.75. De los que compran pintura de látex, 60 % también compra rodillos. Sin embargo, sólo 30 % de los que compran pintura semiesmaltada compra rodillos. Un comprador que se selecciona al azar adquiere un rodillo y una lata de pintura. ¿Cuál es la probabilidad de que sea pintura de látex?

2.102 Denote como *A, B* y *C* a los eventos de que un gran premio se encuentra detrás de las puertas *A, B* y *C,* respectivamente. Suponga que elige al azar una puerta, por ejemplo la *A*. El presentador del juego abre una puerta, por ejemplo la *B*, y muestra que no hay un premio detrás de ella. Ahora, el presentador le da la opción de conservar la puerta que eligió (*A*) o de cambiarla por la puerta que queda (*C*). Utilice la probabilidad para explicar si debe o no hacer el cambio.

Ejercicios de repaso

2.103 Un suero de la verdad tiene la propiedad de que 90% de los sospechosos culpables se juzgan de forma adecuada, mientras que, por supuesto, 10% de los sospechosos culpables erróneamente se consideran inocentes. Por otro lado, a los sospechosos inocentes se les juzga de manera errónea 1% de las veces. Si se aplica el suero a un sospechoso, que se selecciona de un grupo de sospechosos en el cual sólo 5% ha cometido un delito, y éste indica que es culpable, ¿cuál es la probabilidad de que sea inocente?

2.104 Un alergólogo afirma que 50% de los pacientes que examina son alérgicos a algún tipo de hierba. ¿Cuál es la probabilidad de que…

a) exactamente 3 de sus 4 pacientes siguientes sean alérgicos a hierbas?

b) ninguno de sus 4 pacientes siguientes sea alérgico a hierbas?

2.105 Mediante la comparación de las regiones apropiadas en un diagrama de Venn, verifique que

a) $(A \cap B) \cup (A \cap B') = A$;

b) $A' \cap (B' \cup C) = (A' \cap B') \cup (A' \cap C)$.

2.106 Las probabilidades de que una estación de servicio bombee gasolina en 0, 1, 2, 3, 4, 5 o más automóviles durante cierto periodo de 30 minutos son, respectivamente, 0.03, 0.18, 0.24, 0.28, 0.10 y 0.17. Calcule la probabilidad de que en este periodo de 30 minutos

a) más de 2 automóviles reciban gasolina;

b) a lo sumo 4 automóviles reciban gasolina;

c) 4 o más automóviles reciban gasolina.

2.107 ¿Cuántas manos de *bridge* que contengan 4 espadas, 6 diamantes, 1 trébol y 2 corazones son posibles?

2.108 Si la probabilidad de que una persona cometa un error en su declaración de impuestos sobre la renta es 0.1, calcule la probabilidad de que

a) cada una de cuatro personas no relacionadas cometa un error;

b) el señor Jones y la señora Clark cometan un error, y el señor Roberts y la señora Williams no cometan errores.

2.109 Una empresa industrial grande usa tres moteles locales para ofrecer hospedaje nocturno a sus clientes. Se sabe por experiencia que a 20% de los clientes se le asigna habitaciones en el Ramada Inn, a 50% en el Sheraton y a 30% en el Lakeview Motor Lodge. Si hay una falla en la plomería en 5% de las habitaciones del Ramada Inn, en 4% de las habitaciones del Sheraton y en 8% de las habitaciones del Lakeview Motor Lodge, ¿cuál es la probabilidad de que…

a) a un cliente se le asigne una habitación en la que falle la plomería?

b) a una persona que ocupa una habitación en la que falla la plomería se le haya hospedado en el Lakeview Motor Lodge?

2.110 La probabilidad de que un paciente se recupere de una delicada operación de corazón es 0.8. ¿Cuál es la probabilidad de que…

a) exactamente 2 de los siguientes 3 pacientes a los que se somete a esta operación sobrevivan?

b) los siguientes 3 pacientes que tengan esta operación sobrevivan?

2.111 Se sabe que 2/3 de los reclusos en cierta prisión federal son menores de 25 años de edad. También se sabe que 3/5 de los reos son hombres y que 5/8 son mujeres de 25 años de edad o mayores. ¿Cuál es la probabilidad de que un prisionero seleccionado al azar de esta prisión sea mujer y tenga al menos 25 años de edad?

2.112 Si se tienen 4 manzanas rojas, 5 verdes y 6 amarillas, ¿cuántas selecciones de 9 manzanas se pueden hacer si se deben seleccionar 3 de cada color?

2.113 De una caja que contiene 6 bolas negras y 4 verdes se extraen 3 bolas sucesivamente y cada bola se reemplaza en la caja antes de extraer la siguiente. ¿Cuál es la probabilidad de que…

a) las 3 sean del mismo color?

b) cada color esté representado?

2.114 Un cargamento de 12 televisores contiene tres defectuosos. ¿De cuántas formas puede un hotel comprar 5 de estos aparatos y recibir al menos 2 defectuosos?

2.115 Cierto organismo federal emplea a tres empresas consultoras (*A, B* y *C*) con probabilidades de 0.40, 0.35 y 0.25, respectivamente. Se sabe por experiencia que las probabilidades de que las empresas rebasen los costos son 0.05, 0.03 y 0.15, respectivamente. Suponga que el organismo experimenta un exceso en los costos.

a) ¿Cuál es la probabilidad de que la empresa consultora implicada sea la *C*?

b) ¿Cuál es la probabilidad de que sea la *A*?

2.116 Un fabricante estudia los efectos de la temperatura de cocción, el tiempo de cocción y el tipo de aceite para la cocción al elaborar papas fritas. Se utilizan 3 diferentes temperaturas, 4 diferentes tiempos de cocción y 3 diferentes aceites.

a) ¿Cuál es el número total de combinaciones a estudiar?

b) ¿Cuántas combinaciones se utilizarán para cada tipo de aceite?

c) Analice por qué las permutaciones no intervienen en este ejercicio.

2.117 Considere la situación del ejercicio 2.116 y suponga que el fabricante puede probar sólo dos combinaciones en un día.

a) ¿Cuál es la probabilidad de que elija cualquier conjunto dado de 2 corridas?

b) ¿Cuál es la probabilidad de que utilice la temperatura más alta en cualquiera de estas 2 combinaciones?

2.118 Se sabe que existe una probabilidad de 0.07 de que las mujeres de más de 60 años desarrollen cierta forma de cáncer. Se dispone de una prueba de sangre que, aunque no es infalible, permite detectar la enfermedad. De hecho, se sabe que 10 % de las veces la prueba da un falso negativo (es decir, la prueba da un resultado negativo de manera incorrecta) y 5 % de las veces la prueba da un falso positivo (es decir, la prueba da un resultado positivo de manera incorrecta). Si una mujer de más de 60 años se somete a la prueba y recibe un resultado favorable (es decir, negativo), ¿qué probabilidad hay de que tenga la enfermedad?

2.119 Un fabricante de cierto tipo de componente electrónico abastece a los proveedores en lotes de 20. Suponga que 60% de todos los lotes no contiene componentes defectuosos, que 30% contiene un componente defectuoso y que 10% contiene dos componentes defectuosos. Si se elige un lote del que se extraen aleatoriamente dos componentes, los cuales se prueban y ninguno resulta defectuoso,

a) ¿Cuál es la probabilidad de que haya cero componentes defectuosos en el lote?

b) ¿Cuál es la probabilidad de que haya un componente defectuoso en el lote?

c) ¿Cuál es la probabilidad de que haya dos componentes defectuosos en el lote?

2.120 Existe una extraña enfermedad que sólo afecta a uno de cada 500 individuos. Se dispone de una prueba para detectarla, pero, por supuesto, ésta no es infalible. Un resultado correcto positivo (un paciente que realmente tiene la enfermedad) ocurre 95% de las veces; en tanto que un resultado falso positivo (un paciente que no tiene la enfermedad) ocurre 1% de las veces. Si un individuo elegido al azar se somete a prueba y se obtiene un resultado positivo, ¿cuál es la probabilidad de que realmente tenga la enfermedad?

2.121 Una empresa constructora emplea a dos ingenieros de ventas. El ingeniero 1 hace el trabajo de estimar costos en 70% de las cotizaciones solicitadas a la empresa. El ingeniero 2 hace lo mismo en 30% de las

cotizaciones. Se sabe que la tasa de error para el ingeniero 1 es tal que la probabilidad de encontrar un error en su trabajo es 0.02; mientras que la probabilidad de encontrar un error en el trabajo del ingeniero 2 es 0.04. Suponga que al revisar una solicitud de cotización se encuentra un error grave en la estimación de los costos. ¿Qué ingeniero supondría usted que hizo los cálculos? Explique su respuesta y muestre todo el desarrollo.

2.122 En el campo del control de calidad a menudo se usa la ciencia estadística para determinar si un proceso está "fuera de control". Suponga que el proceso, de hecho, está fuera de control y que 20 por ciento de los artículos producidos tiene defecto.

a) Si tres artículos salen en serie de la línea de producción, ¿cuál es la probabilidad de que los tres estén defectuosos?

b) Si salen cuatro artículos en serie, ¿cuál es la probabilidad de que tres estén defectuosos?

2.123 En una planta industrial se está realizando un estudio para determinar la rapidez con la que los trabajadores lesionados regresan a sus labores después del percance. Los registros demuestran que 10% de los trabajadores lesionados son llevados al hospital para su tratamiento y que 15% regresan a su trabajo al día siguiente. Además, los estudios demuestran que 2% son llevados al hospital y regresan al trabajo al día siguiente. Si un trabajador se lesiona, ¿cuál es la probabilidad de que sea llevado al hospital, de que regrese al trabajo al día siguiente, o de ambas cosas?

2.124 Una empresa acostumbra capacitar operadores que realizan ciertas actividades en la línea de producción. Se sabe que los operadores que asisten al curso de capacitación son capaces de cumplir sus cuotas de producción 90% de las veces. Los nuevos operarios que no toman el curso de capacitación sólo cumplen con sus cuotas 65% de las veces. Cincuenta por ciento de los nuevos operadores asisten al curso. Dado que un nuevo operador cumple con su cuota de producción, ¿cuál es la probabilidad de que haya asistido al curso?

2.125 Una encuesta aplicada a quienes usan un software estadístico específico indica que 10% no quedó satisfecho. La mitad de quienes no quedaron satisfechos le compraron el sistema al vendedor *A*. También se sabe que 20% de los encuestados se lo compraron al vendedor *A*. Dado que el proveedor del paquete de software fue el vendedor *A*, ¿cuál es la probabilidad de que un usuario específico haya quedado insatisfecho?

2.126 Durante las crisis económicas se despide a obreros y a menudo se les reemplaza con máquinas. Se revisa la historia de 100 trabajadores cuya pérdida del empleo se atribuye a los avances tecnológicos. Para cada uno de ellos se determinó si obtuvieron un empleo alternativo dentro de la misma empresa, si encontraron un empleo en la misma área de otra empresa, si encontraron trabajo en una nueva área o si llevan desempleados más de un año. Además, se registró la situación sindical de cada trabajador. La siguiente tabla resume los resultados.

	Sindicalizado	No sindicalizado
Está en la misma empresa	40	15
Está en otra empresa (misma área)	13	10
Está en una nueva área	4	11
Está desempleado	2	5

a) Si un trabajador seleccionado encontró empleo en la misma área de una nueva empresa, ¿cuál es la probabilidad de que sea miembro de un sindicato?

b) Si el trabajador es miembro de un sindicato, ¿cuál es la probabilidad de que esté desempleado desde hace un año?

2.127 Hay 50% de probabilidad de que la reina tenga el gen de la hemofilia. Si es portadora, entonces cada uno de los príncipes tiene 50% de probabilidad independiente de tener hemofilia. Si la reina no es portadora, el príncipe no tendrá la enfermedad. Suponga que la reina tuvo tres príncipes que no padecen la enfermedad, ¿cuál es la probabilidad de que la reina sea portadora del gen?

2.128 **Proyecto de equipo:** Entregue a cada estudiante una bolsa de chocolates M&M y forme equipos de 5 o 6 estudiantes. Calcule la distribución de frecuencia relativa del color de los M&M para cada equipo.

a) ¿Cuál es su probabilidad estimada de elegir un chocolate amarillo al azar? ¿Y uno rojo?

b) Ahora haga el mismo cálculo para todo el grupo. ¿Cambiaron las estimaciones?

c) ¿Cree que en un lote procesado existe el mismo número de chocolates de cada color? Comente al respecto.

2.8 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Este capítulo incluye las definiciones, reglas y teoremas fundamentales que convierten a la probabilidad en una herramienta importante para la evaluación de sistemas científicos y de ingeniería. A menudo estas evaluaciones toman la forma de cálculos de probabili-

dad, como se ilustra en los ejemplos y en los ejercicios. Conceptos como independencia, probabilidad condicional, regla de Bayes y otros suelen ser muy adecuados para resolver problemas prácticos en los que se busca obtener un valor de probabilidad. Abundan las ilustraciones en los ejercicios. Vea, por ejemplo, los ejercicios 2.100 y 2.101. En éstos y en muchos otros ejercicios se realiza una evaluación juiciosa de un sistema científico, a partir de un cálculo de probabilidad, utilizando las reglas y las definiciones que se estudian en el capítulo.

Ahora bien, ¿qué relación existe entre el material de este capítulo y el material de otros capítulos? La mejor forma de responder esta pregunta es dando un vistazo al capítulo 3, ya que en éste también se abordan problemas en los que es importante el cálculo de probabilidades. Ahí se ilustra cómo el desempeño de un sistema depende del valor de una o más probabilidades. De nuevo, la probabilidad condicional y la independencia desempeñan un papel. Sin embargo, surgen nuevos conceptos que permiten tener una mayor estructura basada en el concepto de una variable aleatoria y su distribución de probabilidad. Recuerde que el concepto de las distribuciones de frecuencias se abordó brevemente en el capítulo 1. La distribución de probabilidad muestra, en forma gráfica o en una ecuación, toda la información necesaria para describir una estructura de probabilidad. Por ejemplo, en el ejercicio de repaso 2.122 la variable aleatoria de interés es el número de artículos defectuosos, una medición discreta. Por consiguiente, la distribución de probabilidad revelaría la estructura de probabilidad para el número de artículos defectuosos extraídos del número elegido del proceso. Cuando el lector avance al capítulo 3 y los siguientes, será evidente para él que se requieren suposiciones para determinar y, por lo tanto, utilizar las distribuciones de probabilidad en la resolución de problemas científicos.

CAPÍTULO 3

Variables aleatorias y distribuciones de probabilidad

3.1 Concepto de variable aleatoria

La estadística realiza inferencias acerca de las poblaciones y sus características. Se llevan a cabo experimentos cuyos resultados se encuentran sujetos al azar. La prueba de un número de componentes electrónicos es un ejemplo de **experimento estadístico**, un concepto que se utiliza para describir cualquier proceso mediante el cual se generan varias observaciones al azar. A menudo es importante asignar una descripción numérica al resultado. Por ejemplo, cuando se prueban tres componentes electrónicos, el espacio muestral que ofrece una descripción detallada de cada posible resultado se escribe como

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\},$$

donde N denota "no defectuoso", y D, "defectuoso". Es evidente que nos interesa el número de componentes defectuosos que se presenten. De esta forma, a cada punto en el espacio muestral se le *asignará un valor numérico* de 0, 1, 2 o 3. Estos valores son, por supuesto, cantidades aleatorias *determinadas por el resultado del experimento*. Se pueden ver como valores que toma la *variable aleatoria* X, es decir, el número de artículos defectuosos cuando se prueban tres componentes electrónicos.

Definición 3.1: Una **variable aleatoria** es una función que asocia un número real con cada elemento del espacio muestral.

Utilizaremos una letra mayúscula, digamos X, para denotar una variable aleatoria, y su correspondiente letra minúscula, x en este caso, para uno de sus valores. En el ejemplo de la prueba de componentes electrónicos observamos que la variable aleatoria X toma el valor 2 para todos los elementos en el subconjunto

$$E = \{DDN, DND, NDD\}$$

del espacio muestral S. Esto es, cada valor posible de X representa un evento que es un subconjunto del espacio muestral para el experimento dado.

Ejemplo 3.1: De una urna que contiene 4 bolas rojas y 3 negras se sacan 2 bolas de manera sucesiva, sin reemplazo. Los posibles resultados y los valores y de la variable aleatoria Y, donde Y es el número de bolas rojas, son

Espacio muestral	y
RR	2
RN	1
NR	1
NN	0

Ejemplo 3.2: El empleado de un almacén regresa tres cascos de seguridad al azar a tres obreros de un taller siderúrgico que ya los habían probado. Si Smith, Jones y Brown, en ese orden, reciben uno de los tres cascos, liste los puntos muestrales para los posibles órdenes en que el empleado del almacén regresa los cascos, después calcule el valor m de la variable aleatoria M que representa el número de emparejamientos correctos.

***Solución*:** Si S, J y B representan, respectivamente, los cascos que recibieron Smith, Jones y Brown, entonces los posibles arreglos en los cuales se pueden regresar los cascos y el número de emparejamientos correctos son

Espacio muestral	m
SJB	3
SBJ	1
BJS	1
JSB	1
JBS	0
BSJ	0

En cada uno de los dos ejemplos anteriores, el espacio muestral contiene un número finito de elementos. Por el contrario, cuando lanzamos un dado hasta que salga un 5, obtenemos un espacio muestral con una secuencia de elementos interminable,

$$S = \{F, NF, NNF, NNNF, \ldots\},$$

donde F y N representan, respectivamente, la ocurrencia y la no ocurrencia de un 5. Sin embargo, incluso en este experimento el número de elementos se puede igualar a la cantidad total de números enteros, de manera que hay un primer elemento, un segundo, un tercero y así sucesivamente, por lo que se pueden contar.

Hay casos en que la variable aleatoria es categórica por naturaleza en los cuales se utilizan las llamadas variables *ficticias*. Un buen ejemplo de ello es el caso en que la variable aleatoria es binaria por naturaleza, como se indica a continuación.

Ejemplo 3.3: Considere la condición en que salen componentes de la línea de ensamble y se les clasifica como defectuosos o no defectuosos. Defina la variable aleatoria X mediante

$$X = \begin{cases} 1, & \text{si el componente está defectuoso,} \\ 0, & \text{si el componente no está defectuoso.} \end{cases}$$

Evidentemente la asignación de 1 o 0 es arbitraria, aunque bastante conveniente, lo cual quedará más claro en capítulos posteriores. La variable aleatoria en la que se eligen 0 y 1 para describir los dos posibles valores se denomina **variable aleatoria de Bernoulli**.

En los siguientes ejemplos veremos más casos de variables aleatorias.

Ejemplo 3.4: Los estadísticos utilizan **planes de muestreo** para aceptar o rechazar lotes de materiales. Suponga que uno de los planes de muestreo implica obtener una muestra independiente de 10 artículos de un lote de 100, en el que 12 están defectuosos.

Si X representa a la variable aleatoria, definida como el número de artículos que están defectuosos en la muestra de 10, la variable aleatoria toma los valores 0, 1, 2, . . . , 9, 10.

Ejemplo 3.5: Suponga que un plan de muestreo implica obtener una muestra de artículos de un proceso hasta que se encuentre uno defectuoso. La evaluación del proceso dependerá de cuántos artículos consecutivos se observen. En este caso, sea X una variable aleatoria que se define como el número de artículos observados antes de que salga uno defectuoso. Si N representa un artículo no defectuoso y D uno defectuoso, los espacios muestrales son $S = \{D\}$ dado que $X = 1$, $S = \{ND\}$ dado que $X = 2$, $S = \{NND\}$ dado que $X = 3$, y así sucesivamente.

Ejemplo 3.6: Existe interés por la proporción de personas que responden a cierta encuesta enviada por correo. Sea X tal proporción. X es una variable aleatoria que toma todos los valores de x para los cuales $0 \leq x \leq 1$.

Ejemplo 3.7: Sea X la variable aleatoria definida como el tiempo que pasa, en horas, para que un radar detecte entre conductores sucesivos a los que exceden los límites de velocidad. La variable aleatoria X toma todos los valores de x para los que $x \geq 0$.

Definición 3.2: Si un espacio muestral contiene un número finito de posibilidades, o una serie interminable con tantos elementos como números enteros existen, se llama **espacio muestral discreto**.

Los resultados de algunos experimentos estadísticos no pueden ser ni finitos ni contables. Éste es el caso, por ejemplo, en una investigación que se realiza para medir las distancias que recorre un automóvil de cierta marca, en una ruta de prueba preestablecida, con cinco litros de gasolina. Si se asume que la distancia es una variable que se mide con algún grado de precisión, entonces salta a la vista que tenemos un número infinito de distancias posibles en el espacio muestral, que no se pueden igualar a la cantidad total de números enteros. Lo mismo ocurre en el caso de un experimento en que se registra el tiempo requerido para que ocurra una reacción química, en donde una vez más los posibles intervalos de tiempo que forman el espacio muestral serían un número infinito e incontable. Vemos ahora que no todos los espacios muestrales necesitan ser discretos.

Definición 3.3: Si un espacio muestral contiene un número infinito de posibilidades, igual al número de puntos en un segmento de recta, se le denomina **espacio muestral continuo**.

Una variable aleatoria se llama **variable aleatoria discreta** si se puede contar su conjunto de resultados posibles. En los ejemplos 3.1 a 3.5 las variables aleatorias son discretas. Sin embargo, una variable aleatoria cuyo conjunto de valores posibles es un intervalo completo de números no es discreta. Cuando una variable aleatoria puede tomar valores

en una escala continua, se le denomina **variable aleatoria continua**. A menudo los posibles valores de una variable aleatoria continua son precisamente los mismos valores incluidos en el espacio muestral continuo. Es evidente que las variables aleatorias descritas en los ejemplos 3.6 y 3.7 son variables aleatorias continuas.

En la mayoría de los problemas prácticos las variables aleatorias continuas representan datos *medidos*, como serían todos los posibles pesos, alturas, temperaturas, distancias o periodos de vida; en tanto que las variables aleatorias discretas representan datos *por conteo*, como el número de artículos defectuosos en una muestra de k artículos o el número de accidentes de carretera por año en una entidad específica. Observe que tanto Y como M, las variables aleatorias de los ejemplos 3.1 y 3.2, representan datos por conteo: Y el número de bolas rojas y M el número de emparejamientos correctos de cascos.

3.2 Distribuciones discretas de probabilidad

Una variable aleatoria discreta toma cada uno de sus valores con cierta probabilidad. Al lanzar una moneda tres veces, la variable X, que representa el número de caras, toma el valor 2 con $3/8$ de probabilidad, pues 3 de los 8 puntos muestrales igualmente probables tienen como resultado dos caras y una cruz. Si se suponen pesos iguales para los eventos simples del ejemplo 3.2, la probabilidad de que ningún obrero reciba el casco correcto, es decir, la probabilidad de que M tome el valor cero, es $1/3$. Los valores posibles m de M y sus probabilidades son

m	0	1	3
$P(M = m)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Observe que los valores de m agotan todos los casos posibles, por lo tanto, las probabilidades suman 1.

Con frecuencia es conveniente representar todas las probabilidades de una variable aleatoria X usando una fórmula, la cual necesariamente sería una función de los valores numéricos x que denotaremos con $f(x)$, $g(x)$, $r(x)$ y así sucesivamente. Por lo tanto, escribimos $f(x) = P(X = x)$; es decir, $f(3) = P(X = 3)$. Al conjunto de pares ordenados $(x, f(x))$ se le llama **función de probabilidad**, **función de masa de probabilidad** o **distribución de probabilidad** de la variable aleatoria discreta X.

Definición 3.4: El conjunto de pares ordenados $(x, f(x))$ es una **función de probabilidad**, una **función de masa de probabilidad** o una **distribución de probabilidad** de la variable aleatoria discreta X si, para cada resultado x posible,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Ejemplo 3.8: Un embarque de 20 computadoras portátiles similares para una tienda minorista contiene 3 que están defectuosas. Si una escuela compra al azar 2 de estas computadoras, calcule la distribución de probabilidad para el número de computadoras defectuosas.

***Solución*:** Sea X una variable aleatoria cuyos valores x son los números posibles de computadoras defectuosas compradas por la escuela. Entonces x sólo puede asumir los números 0, 1 y 2. Así,

$$f(0) = P(X = 0) = \frac{\binom{3}{0}\binom{17}{2}}{\binom{20}{2}} = \frac{68}{95}, \quad f(1) = P(X = 1) = \frac{\binom{3}{1}\binom{17}{1}}{\binom{20}{2}} = \frac{51}{190},$$
$$f(2) = P(X = 2) = \frac{\binom{3}{2}\binom{17}{0}}{\binom{20}{2}} = \frac{3}{190}.$$

Por consiguiente, la distribución de probabilidad de X es

x	0	1	2
$f(x)$	$\frac{68}{95}$	$\frac{51}{190}$	$\frac{3}{190}$

Ejemplo 3.9: Si una agencia automotriz vende 50% de su inventario de cierto vehículo extranjero equipado con bolsas de aire laterales, calcule una fórmula para la distribución de probabilidad del número de automóviles con bolsas de aire laterales entre los siguientes 4 vehículos que venda la agencia.

***Solución*:** Como la probabilidad de vender un automóvil con bolsas de aire laterales es 0.5, los $2^4 = 16$ puntos del espacio muestral tienen la misma probabilidad de ocurrencia. Por lo tanto, el denominador para todas las probabilidades, y también para nuestra función, es 16. Para obtener el número de formas de vender tres automóviles con bolsas de aire laterales necesitamos considerar el número de formas de dividir 4 resultados en 2 celdas, con 3 automóviles con bolsas de aire laterales asignados a una celda, y el modelo sin bolsas de aire laterales asignado a la otra. Esto se puede hacer de $\binom{4}{3} = 4$ formas. En general, el evento de vender x modelos con bolsas de aire laterales y $4 - x$ modelos sin bolsas de aire laterales puede ocurrir de $\binom{4}{x}$ formas, donde x puede ser 0, 1, 2, 3 o 4. Por consiguiente, la distribución de probabilidad $f(x) = P(X = x)$ es

$$f(x) = \frac{1}{16}\binom{4}{x}, \quad \text{para } x = 0, 1, 2, 3, 4.$$

Existen muchos problemas en los que desearíamos calcular la probabilidad de que el valor observado de una variable aleatoria X sea menor o igual que algún número real x. Al escribir $F(x) = P(X \leq x)$ para cualquier número real x, definimos $F(x)$ como la **función de la distribución acumulativa** de la variable aleatoria X.

Definición 3.5: La **función de la distribución acumulativa** $F(x)$ de una variable aleatoria discreta X con distribución de probabilidad $f(x)$ es

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{para} \quad -\infty < x < \infty.$$

Para la variable aleatoria M, el número de emparejamientos correctos en el ejemplo 3.2, tenemos

$$F(2) = P(M \leq 2) = f(0) + f(1) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}.$$

La función de la distribución acumulativa de M es

$$F(m) = \begin{cases} 0, & \text{para } m < 0, \\ \frac{1}{3}, & \text{para } 0 \leq m < 1, \\ \frac{5}{6}, & \text{para } 1 \leq m < 3, \\ 1, & \text{para } m \geq 3. \end{cases}$$

Es necesario observar en particular el hecho de que la función de la distribución acumulativa es una función no decreciente monótona, la cual no sólo se define para los valores que toma la variable aleatoria dada sino para todos los números reales.

Ejemplo 3.10: Calcule la función de la distribución acumulativa de la variable aleatoria X del ejemplo 3.9. Utilice $F(x)$ para verificar que $f(2) = 3/8$.

***Solución*:** El cálculo directo de la distribución de probabilidad del ejemplo 3.9 da $f(0) = 1/16$, $f(1) = 1/4, f(2) = 3/8, f(3) = 1/4$ y $f(4) = 1/16$. Por lo tanto,

$$\begin{aligned}
F(0) &= f(0) = \frac{1}{16},\\
F(1) &= f(0) + f(1) = \frac{5}{16},\\
F(2) &= f(0) + f(1) + f(2) = \frac{11}{16},\\
F(3) &= f(0) + f(1) + f(2) + f(3) = \frac{15}{16},\\
F(4) &= f(0) + f(1) + f(2) + f(3) + f(4) = 1.
\end{aligned}$$

Por lo tanto,

$$F(x) = \begin{cases} 0, & \text{para } x < 0,\\ \frac{1}{16}, & \text{para } 0 \le x < 1,\\ \frac{5}{16}, & \text{para } 1 \le x < 2,\\ \frac{11}{16}, & \text{para } 2 \le x < 3,\\ \frac{15}{16}, & \text{para } 3 \le x < 4,\\ 1 & \text{para } x \ge 4. \end{cases}$$

Entonces,

$$f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}.$$

A menudo es útil ver una distribución de probabilidad en forma gráfica. Se pueden graficar los puntos $(x, f(x))$ del ejemplo 3.9 para obtener la figura 3.1. Si unimos los puntos al eje x, ya sea con una línea punteada o con una línea sólida, obtenemos una gráfica de función de masa de probabilidad. La figura 3.1 permite ver fácilmente qué valores de X tienen más probabilidad de ocurrencia y, en este caso, también indica una situación perfectamente simétrica.

Sin embargo, en vez de graficar los puntos $(x, f(x))$, lo que hacemos más a menudo es construir rectángulos como en la figura 3.2. Aquí los rectángulos se construyen de manera que sus bases, con la misma anchura, se centren en cada valor x, y que sus alturas igualen a las probabilidades correspondientes dadas por $f(x)$. Las bases se construyen de forma tal que no dejen espacios entre los rectángulos. La figura 3.2 se denomina **histograma de probabilidad**.

Como cada base en la figura 3.2 tiene el ancho de una unidad, $P(X = x)$ es igual al área del rectángulo centrado en x. Incluso si las bases no tuvieran el ancho de una unidad, podríamos ajustar las alturas de los rectángulos para que tengan áreas que igualen las probabilidades de X de tomar cualquiera de sus valores x. Este concepto de utilizar

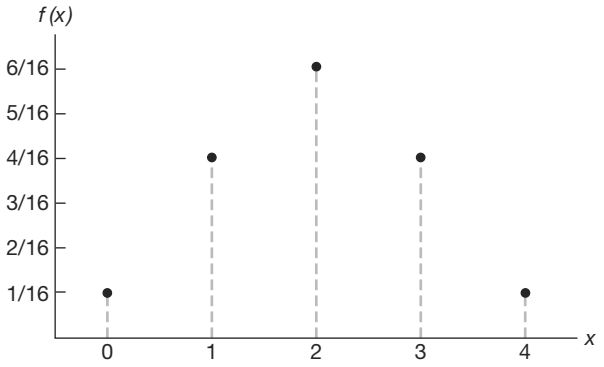


Figura 3.1: Gráfica de función de masa de probabilidad.

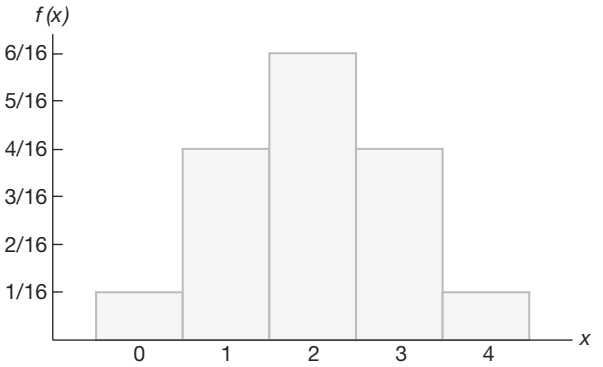


Figura 3.2: Histograma de probabilidad.

áreas para representar probabilidades es necesario para nuestro estudio de la distribución de probabilidad de una variable aleatoria continua.

La gráfica de la función de la distribución acumulativa del ejemplo 3.9, que aparece como una función escalonada en la figura 3.3, se obtiene graficando los puntos $(x, F(x))$.

Ciertas distribuciones de probabilidad se aplican a más de una situación física. La distribución de probabilidad del ejemplo 3.9 también se aplica a la variable aleatoria Y, donde Y es el número de caras que se obtienen cuando una moneda se lanza 4 veces, o a la variable aleatoria W, donde W es el número de cartas rojas que resultan cuando se sacan 4 cartas al azar de una baraja de manera sucesiva, se reemplaza cada carta y se baraja antes de sacar la siguiente. En el capítulo 5 se estudiarán distribuciones discretas especiales que se aplican a diversas situaciones experimentales.

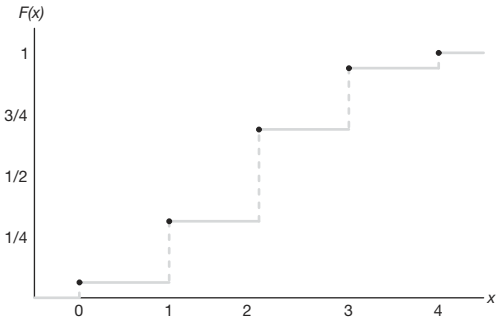


Figura 3.3: Función de distribución acumulativa discreta.

3.3 Distribuciones de probabilidad continua

Una variable aleatoria continua tiene una probabilidad 0 de adoptar *exactamente* cualquiera de sus valores. En consecuencia, su distribución de probabilidad no se puede

presentar en forma tabular. En un principio esto parecería sorprendente, pero se vuelve más probable cuando consideramos un ejemplo específico. Consideremos una variable aleatoria cuyos valores son las estaturas de todas las personas mayores de 21 años de edad. Entre cualesquiera dos valores, digamos 163.5 y 164.5 centímetros, o incluso entre 163.99 y 164.01 centímetros, hay un número infinito de estaturas, una de las cuales es 164 centímetros. La probabilidad de seleccionar al azar a una persona que tenga exactamente 164 centímetros de estatura en lugar de una del conjunto infinitamente grande de estaturas tan cercanas a 164 centímetros que humanamente no sea posible medir la diferencia es remota, por consiguiente, asignamos una probabilidad 0 a tal evento. Sin embargo, esto no ocurre si nos referimos a la probabilidad de seleccionar a una persona que mida al menos 163 centímetros pero no más de 165 centímetros de estatura. Aquí nos referimos a un intervalo en vez de a un valor puntual de nuestra variable aleatoria.

Nos interesamos por el cálculo de probabilidades para varios intervalos de variables aleatorias continuas como $P(a < X < b)$, $P(W \geq c)$, etc. Observe que cuando X es continua,

$$P(a < X \leq b) = P(a < X < b) + P(X = b) = P(a < X < b).$$

Es decir, no importa si incluimos o no un extremo del intervalo. Sin embargo, esto no es cierto cuando X es discreta.

Aunque la distribución de probabilidad de una variable aleatoria continua no se puede representar de forma tabular, sí es posible plantearla como una fórmula, la cual necesariamente será función de los valores numéricos de la variable aleatoria continua X, y como tal se representará mediante la notación funcional $f(x)$. Cuando se trata con variables continuas, a $f(x)$ por lo general se le llama **función de densidad de probabilidad**, o simplemente **función de densidad** de X. Como X se define sobre un espacio muestral continuo, es posible que $f(x)$ tenga un número finito de discontinuidades. Sin embargo, la mayoría de las funciones de densidad que tienen aplicaciones prácticas en el análisis de datos estadísticos son continuas y sus gráficas pueden tomar cualesquiera de varias formas, algunas de las cuales se presentan en la figura 3.4. Como se utilizarán áreas para representar probabilidades y éstas son valores numéricos positivos, la función de densidad debe caer completamente arriba del eje x.

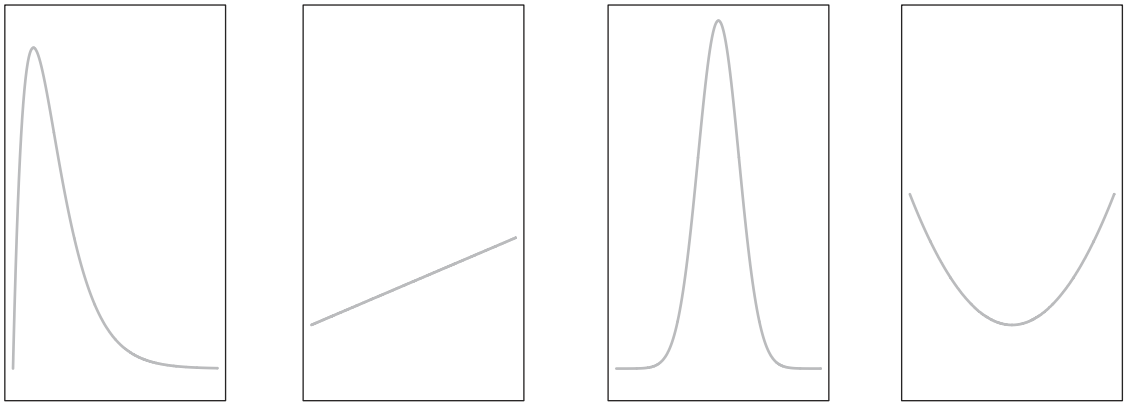


Figura 3.4: Funciones de densidad típicas.

Una función de densidad de probabilidad se construye de manera que el área bajo su curva limitada por el eje x sea igual a 1, cuando se calcula en el rango de X para el que se define $f(x)$. Como este rango de X es un intervalo finito, siempre es posible extender el intervalo para que incluya a todo el conjunto de números reales definiendo $f(x)$ como cero en todos los puntos de las partes extendidas del intervalo. En la figura 3.5 la probabilidad de que X tome un valor entre a y b es igual al área sombreada bajo la función de densidad entre las ordenadas en $x = a$ y $x = b$, y a partir del cálculo integral está dada por

$$P(a < X < b) = \int_a^b f(x)\, dx.$$

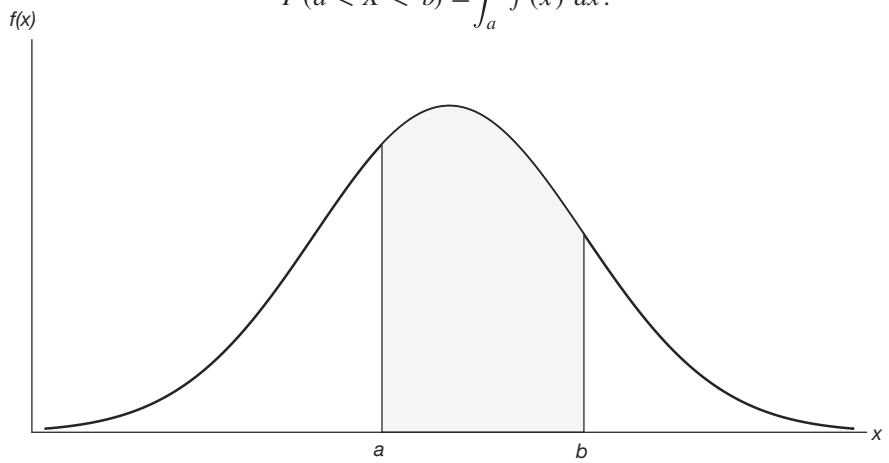


Figura 3.5: $P(a < X < b)$.

Definición 3.6: La función $f(x)$ es una **función de densidad de probabilidad** (fdp) para la variable aleatoria continua X, definida en el conjunto de números reales, si

1. $f(x) \geq 0$, para toda $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

Ejemplo 3.11: Suponga que el error en la temperatura de reacción, en °C, en un experimento de laboratorio controlado, es una variable aleatoria continua X que tiene la función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique que $f(x)$ es una función de densidad.

b) Calcule $P(0 < X \leq 1)$.

***Solución*:** Usamos la definición 3.6.

a) Evidentemente, $f(x) \geq 0$. Para verificar la condición 2 de la definición 3.6 tenemos

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}\Big|_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1.$$

b) Si usamos la fórmula 3 de la definición 3.6, obtenemos

$$P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3} dx = \frac{x^3}{9}\Big|_0^1 = \frac{1}{9}.$$

Definición 3.7: La **función de distribución acumulativa** $F(x)$, de una variable aleatoria continua X con función de densidad $f(x)$, es

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \text{ para } -\infty < x < \infty.$$

Como una consecuencia inmediata de la definición 3.7 se pueden escribir los dos resultados,

$$P(a < X < b) = F(b) - F(a) \text{ y } f(x) = \frac{dF(x)}{dx},$$

si existe la derivada.

Ejemplo 3.12: Calcule $F(x)$ para la función de densidad del ejemplo 3.11 y utilice el resultado para evaluar $P(0 < X \leq 1)$.

***Solución*:** Para $-1 < x < 2$,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-1}^{x} \frac{t^2}{3} dt = \frac{t^3}{9}\Big|_{-1}^{x} = \frac{x^3 + 1}{9}.$$

Por lo tanto,

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

La función de la distribución acumulativa $F(x)$ se expresa en la figura 3.6. Entonces,

$$P(0 < X \leq 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9},$$

que coincide con el resultado que se obtuvo al utilizar la función de densidad en el ejemplo 3.11.

Ejemplo 3.13: El Departamento de Energía (DE) asigna proyectos mediante licitación y, por lo general, estima lo que debería ser una licitación razonable. Sea b el estimado. El DE determinó que la función de densidad de la licitación ganadora (baja) es

$$f(y) = \begin{cases} \frac{5}{8b}, & \frac{2}{5}b \leq y \leq 2b, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule $F(y)$ y utilice el resultado para determinar la probabilidad de que la licitación ganadora sea menor que la estimación preliminar b del DE.

***Solución*:** Para $2b/5 \leq y \leq 2b$,

$$F(y) = \int_{2b/5}^{y} \frac{5}{8b} dy = \frac{5t}{8b}\Big|_{2b/5}^{y} = \frac{5y}{8b} - \frac{1}{4}.$$

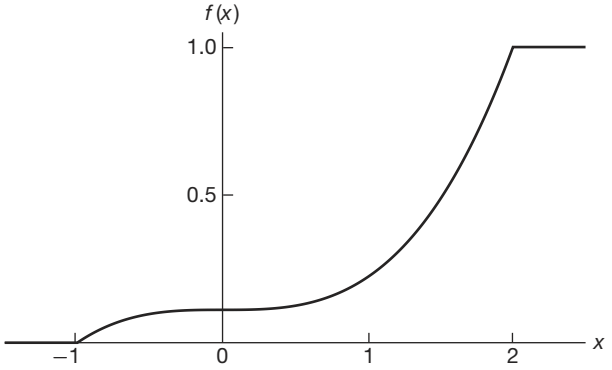


Figura 3.6: Función de distribución acumulativa continua.

Por consiguiente,

$$F(y) = \begin{cases} 0, & y < \frac{2}{5}b, \\ \frac{5y}{8b} - \frac{1}{4}, & \frac{2}{5}b \le y < 2b, \\ 1, & y \ge 2b. \end{cases}$$

Para determinar la probabilidad de que la licitación ganadora sea menor que la estimación preliminar b de la licitación tenemos

$$P(Y \le b) = F(b) = \frac{5}{8} - \frac{1}{4} = \frac{3}{8}.$$

Ejercicios

3.1 Clasifique las siguientes variables aleatorias como discretas o continuas:

X: el número de accidentes automovilísticos que ocurren al año en Virginia.

Y: el tiempo para jugar 18 hoyos de golf.

M: la cantidad de leche que una vaca específica produce anualmente.

N: el número de huevos que una gallina pone mensualmente.

P: el número de permisos para construcción que los funcionarios de una ciudad emiten cada mes.

Q: el peso del grano producido por acre.

3.2 Un embarque foráneo de 5 automóviles extranjeros contiene 2 que tienen ligeras manchas de pintura. Suponga que una agencia recibe 3 de estos automóviles al azar y liste los elementos del espacio muestral S usando las letras M y N para "manchado" y "sin mancha", respectivamente; luego asigne a cada punto muestral un valor x de la variable aleatoria X que representa el número de automóviles con manchas de pintura que compró la agencia.

3.3 Sea W la variable aleatoria que da el número de caras menos el número de cruces en tres lanzamientos de una moneda. Liste los elementos del espacio muestral S para los tres lanzamientos de la moneda y asigne un valor w de W a cada punto muestral.

3.4 Se lanza una moneda hasta que se presentan 3 caras sucesivamente. Liste sólo aquellos elementos del espacio muestral que requieren 6 o menos lanzamientos. ¿Es éste un espacio muestral discreto? Explique su respuesta.

3.5 Determine el valor c de modo que cada una de las siguientes funciones sirva como distribución de probabilidad de la variable aleatoria discreta X:

a) $f(x) = c(x^2 + 4)$, para $x = 0, 1, 2, 3$;

b) $f(x) = c\binom{2}{x}\binom{3}{3-x}$, para $x = 0, 1, 2$.

3.6 La vida útil, en días, para frascos de cierta medicina de prescripción es una variable aleatoria que tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{20,000}{(x+100)^3}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la probabilidad de que un frasco de esta medicina tenga una vida útil de
a) al menos 200 días;
b) cualquier lapso entre 80 y 120 días.

3.7 El número total de horas, medidas en unidades de 100 horas, que una familia utiliza una aspiradora en un periodo de un año es una variable aleatoria continua X que tiene la siguiente función de densidad:

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la probabilidad de que en un periodo de un año una familia utilice su aspiradora
a) menos de 120 horas;
b) entre 50 y 100 horas.

3.8 Obtenga la distribución de probabilidad de la variable aleatoria W del ejercicio 3.3; suponga que la moneda está cargada, de manera que existe el doble de probabilidad de que ocurra una cara que una cruz.

3.9 La proporción de personas que responden a cierta encuesta enviada por correo es una variable aleatoria continua X que tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{2(x+2)}{5}, & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Demuestre que $P(0 < X < 1) = 1$.
b) Calcule la probabilidad de que más de $1/4$ pero menos de $1/2$ de las personas contactadas respondan a este tipo de encuesta.

3.10 Encuentre una fórmula para la distribución de probabilidad de la variable aleatoria X que represente el resultado cuando se lanza un dado una vez.

3.11 Un embarque de 7 televisores contiene 2 unidades defectuosas. Un hotel compra 3 de los televisores al azar. Si x es el número de unidades defectuosas que compra el hotel, calcule la distribución de probabilidad de X. Exprese los resultados de forma gráfica como un histograma de probabilidad.

3.12 Una empresa de inversiones ofrece a sus clientes bonos municipales que vencen después de varios años. Dado que la función de distribución acumulativa de T, el número de años para el vencimiento de un bono que se elige al azar, es

$$F(t) = \begin{cases} 0, & t < 1, \\ \frac{1}{4}, & 1 \le t < 3, \\ \frac{1}{2}, & 3 \le t < 5, \\ \frac{3}{4}, & 5 \le t < 7, \\ 1, & t \ge 7, \end{cases}$$

calcule
a) $P(T = 5)$;
b) $P(T > 3)$;
c) $P(1.4 < T < 6)$;
d) $P(T \le 5 \mid T \ge 2)$;

3.13 La distribución de probabilidad de X, el número de imperfecciones que se encuentran en cada 10 metros de una tela sintética que viene en rollos continuos de ancho uniforme, está dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Construya la función de distribución acumulativa de X.

3.14 El tiempo que pasa, en horas, para que un radar detecte entre conductores sucesivos a los que exceden los límites de velocidad es una variable aleatoria continua con una función de distribución acumulativa

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-8x}, & x \ge 0. \end{cases}$$

Calcule la probabilidad de que el tiempo que pase para que el radar detecte entre conductores sucesivos a los que exceden los límites de velocidad sea menor de 12 minutos
a) usando la función de distribución acumulativa de X;
b) utilizando la función de densidad de probabilidad de X.

3.15 Calcule la función de distribución acumulativa de la variable aleatoria X que represente el número de unidades defectuosas en el ejercicio 3.11. Luego, utilice $F(x)$ para calcular
a) $P(X = 1)$;
b) $P(0 < X \le 2)$.

3.16 Construya una gráfica de la función de distribución acumulativa del ejercicio 3.15.

3.17 Una variable aleatoria continua X, que puede tomar valores entre $x = 1$ y $x = 3$, tiene una función de densidad dada por $f(x) = 1/2$.
a) Muestre que el área bajo la curva es igual a 1.
b) Calcule $P(2 < X < 2.5)$.
c) Calcule $P(X \le 1.6)$.

3.18 Una variable aleatoria continua *X,* que puede tomar valores entre $x = 2$ y $x = 5$, tiene una función de densidad dada por $f(x) = 2(1 + x)/27$. Calcule
a) $P(X < 4)$;
b) $P(3 \leq X < 4)$.

3.19 Para la función de densidad del ejercicio 3.17 calcule $F(x)$. Utilícela para evaluar $P(2 < X < 2.5)$.

3.20 Para la función de densidad del ejercicio 3.18 calcule $F(x)$ y utilícela para evaluar $P(3 \leq X < 4)$.

3.21 Considere la función de densidad

$$f(x) = \begin{cases} k\sqrt{x}, & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Evalúe k.
b) Calcule $F(x)$ y utilice el resultado para evaluar

$$P(0.3 < X < 0.6).$$

3.22 Se sacan tres cartas de una baraja de manera sucesiva y sin reemplazo. Calcule la distribución de probabilidad para la cantidad de espadas.

3.23 Calcule la función de distribución acumulativa de la variable aleatoria W del ejercicio 3.8. Use $F(w)$ para calcular
a) $P(W > 0)$;
b) $P(-1 \leq W < 3)$.

3.24 Calcule la distribución de probabilidad para el número de discos compactos de jazz cuando, de una colección que consta de 5 de jazz, 2 de música clásica y 3 de rock, se seleccionan 4 CD al azar. Exprese sus resultados utilizando una fórmula.

3.25 De una caja que contiene 4 monedas de 10 centavos y 2 monedas de 5 centavos se seleccionan 3 monedas al azar y sin reemplazo. Calcule la distribución de probabilidad para el total T de las 3 monedas. Exprese la distribución de probabilidad de forma gráfica como un histograma de probabilidad.

3.26 De una caja que contiene 4 bolas negras y 2 verdes se sacan 3 bolas sucesivamente, cada bola se regresa a la caja antes de sacar la siguiente. Calcule la distribución de probabilidad para el número de bolas verdes.

3.27 El tiempo que pasa, en horas, antes de que una parte importante de un equipo electrónico que se utiliza para fabricar un reproductor de DVD empiece a fallar tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{1}{2000} \exp(-x/2000), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

a) Calcule $F(x)$.
b) Determine la probabilidad de que el componente (y, por lo tanto, el reproductor de DVD) funcionedurante más de 1000 horas antes de que sea necesario reemplazar el componente.
c) Determine la probabilidad de que el componente falle antes de 2000 horas.

3.28 Un productor de cereales está consciente de que el peso del producto varía ligeramente entre una y otra caja. De hecho, cuenta con suficientes datos históricos para determinar la función de densidad que describe la estructura de probabilidad para el peso (en onzas). Si X es el peso, en onzas, de la variable aleatoria, la función de densidad se describe como

$$f(x) = \begin{cases} \frac{2}{5}, & 23.75 \leq x \leq 26.25, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique que sea una función de densidad válida.
b) Determine la probabilidad de que el peso sea menor que 24 onzas.
c) La empresa desea que un peso mayor que 26 onzas sea un caso extraordinariamente raro. ¿Cuál será la probabilidad de que en verdad ocurra este caso extraordinariamente raro?

3.29 Un factor importante en el combustible sólido para proyectiles es la distribución del tamaño de las partículas. Cuando las partículas son demasiado grandes se presentan problemas importantes. A partir de datos de producción históricos se determinó que la distribución del tamaño (en micras) de las partículas se caracteriza por

$$f(x) = \begin{cases} 3x^{-4}, & x > 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique que sea una función de densidad válida.
b) Evalúe $F(x)$.
c) ¿Cuál es la probabilidad de que una partícula tomada al azar del combustible fabricado sea mayor que 4 micras?

3.30 Las mediciones en los sistemas científicos siempre están sujetas a variación, algunas veces más que otras. Hay muchas estructuras para los errores de medición y los estadísticos pasan mucho tiempo modelándolos. Suponga que el error de medición X de cierta cantidad física es determinado por la siguiente función de densidad:

$$f(x) = \begin{cases} k(3 - x^2), & -1 \leq x \leq 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine k, que representa $f(x)$, una función de densidad válida.
b) Calcule la probabilidad de que un error aleatorio en la medición sea menor que ½.
c) Para esta medición específica, resulta indeseable si la *magnitud* del error (es decir, $|x|$) es mayor que 0.8. ¿Cuál es la probabilidad de que esto ocurra?

3.31 Con base en pruebas extensas, el fabricante de una lavadora determinó que el tiempo Y (en años) para que el electrodoméstico requiera una reparación mayor se obtiene mediante la siguiente función de densidad de probabilidad:

$$f(y) = \begin{cases} \frac{1}{4}e^{-y/4}, & y \geq 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Los críticos considerarían que la lavadora es una ganga si no hay probabilidades de que requiera una reparación mayor antes del sexto año. Comente sobre esto determinando $P(Y > 6)$.
b) ¿Cuál es la probabilidad de que la lavadora requiera una reparación mayor durante el primer año?

3.32 Se está revisando qué proporciones de su presupuesto asigna cierta empresa industrial a controles ambientales y de contaminación. Un proyecto de recopilación de datos determina que la distribución de tales proporciones está dada por

$$f(y) = \begin{cases} 5(1-y)^4, & 0 \leq y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que la función de densidad anterior sea válida.
b) ¿Cuál es la probabilidad de que una empresa elegida al azar gaste menos de 10% de su presupuesto en controles ambientales y de contaminación?
c) ¿Cuál es la probabilidad de que una empresa seleccionada al azar gaste más de 50% de su presupuesto en controles ambientales y de la contaminación?

3.33 Suponga que cierto tipo de pequeñas empresas de procesamiento de datos están tan especializadas que algunas tienen dificultades para obtener utilidades durante su primer año de operación. La función de densidad de probabilidad que caracteriza la proporción Y que obtiene utilidades está dada por

$$f(y) = \begin{cases} ky^4(1-y)^3, & 0 \leq y \leq 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) ¿Cuál es el valor de k que hace de la anterior una función de densidad válida?
b) Calcule la probabilidad de que al menos 50% de las empresas tenga utilidades durante el primer año.
c) Calcule la probabilidad de que al menos 80% de las empresas tenga utilidades durante el primer año.

3.34 Los tubos de magnetrón se producen en una línea de ensamble automatizada. Periódicamente se utiliza un plan de muestreo para evaluar la calidad en la longitud de los tubos; sin embargo, dicha medida está sujeta a incertidumbre. Se considera que la probabilidad de que un tubo elegido al azar cumpla con las especificaciones de longitud es 0.99. Se utiliza un plan de muestreo en el cual se mide la longitud de 5 tubos elegidos al azar.

a) Muestre que la función de probabilidad de Y, el número de tubos de cada 5 que cumplen con las especificaciones de longitud, está dada por la siguiente función de probabilidad discreta:

$$f(y) = \frac{5!}{y!(5-y)!}(0.99)^y(0.01)^{5-y},$$

b) Suponga que se eligen artículos de la línea al azar y 3 no cumplen con las especificaciones. Utilice la $f(y)$ anterior para apoyar o refutar la conjetura de que hay 0.99 de probabilidades de que un solo tubo cumpla con las especificaciones.

3.35 Suponga que a partir de gran cantidad de datos históricos se sabe que X, el número de automóviles que llegan a una intersección específica durante un periodo de 20 segundos, se determina mediante la siguiente función de probabilidad discreta

$$f(x) = e^{-6}\frac{6^x}{x!}, \quad \text{para} \quad x = 0, 1, 2, \ldots.$$

a) Calcule la probabilidad de que en un periodo específico de 20 segundos más de 8 automóviles lleguen a la intersección.
b) Calcule la probabilidad de que sólo lleguen 2 automóviles.

3.36 En una tarea de laboratorio, si el equipo está funcionando, la función de densidad del resultado observado, X, es

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule $P(X \leq 1/3)$.
b) ¿Cuál es la probabilidad de que X sea mayor que 0.5?
c) Dado que $X \geq 0.5$, ¿cuál es la probabilidad de que X sea menor que 0.75?

3.4 Distribuciones de probabilidad conjunta

El estudio de las variables aleatorias y sus distribuciones de probabilidad de la sección anterior se restringió a espacios muestrales unidimensionales, ya que registramos los resultados de un experimento como los valores que toma una sola variable aleatoria. No

obstante, habrá situaciones en las que se busque registrar los resultados simultáneos de diversas variables aleatorias. Por ejemplo, en un experimento químico controlado podríamos medir la cantidad del precipitado P y la del volumen V de gas liberado, lo que daría lugar a un espacio muestral bidimensional que consta de los resultados (p, v); o bien, podríamos interesarnos en la dureza d y en la resistencia a la tensión T de cobre estirado en frío que produciría los resultados (d, t). En un estudio realizado con estudiantes universitarios para determinar la probabilidad de que tengan éxito en la universidad, basado en los datos del nivel preparatoria, se podría utilizar un espacio muestral tridimensional y registrar la calificación que obtuvo cada uno en la prueba de aptitudes, el lugar que cada uno ocupó en la preparatoria y la calificación promedio que cada uno obtuvo al final de su primer año en la universidad.

Si X y Y son dos variables aleatorias discretas, la distribución de probabilidad para sus ocurrencias simultáneas se representa mediante una función con valores $f(x, y)$, para cualquier par de valores (x, y) dentro del rango de las variables aleatorias X y Y. Se acostumbra referirse a esta función como la **distribución de probabilidad conjunta** de X y Y.

Por consiguiente, en el caso discreto,

$$f(x, y) = P(X = x, Y = y);$$

es decir, los valores $f(x, y)$ dan la probabilidad de que los resultados x y y ocurran al mismo tiempo. Por ejemplo, si se le va a dar servicio a los neumáticos de un camión de transporte pesado, y X representa el número de millas que éstos han recorrido y Y el número de neumáticos que deben ser reemplazados, entonces $f(30{,}000, 5)$ es la probabilidad de que los neumáticos hayan recorrido más de 30,000 millas y que el camión necesite 5 neumáticos nuevos.

Definición 3.8: La función $f(x,y)$ es una **distribución de probabilidad conjunta** o **función de masa de probabilidad** de las variables aleatorias discretas X y Y, si

1. $f(x, y) \geq 0$ para toda (x, y),
2. $\sum_x \sum_y f(x, y) = 1$,
3. $P(X = x, Y = y) = f(x, y)$.

Para cualquier región A en el plano xy, $P[(X, Y) \in A] = \sum\sum_A f(x,y)$.

Ejemplo 3.14: Se seleccionan al azar 2 repuestos para bolígrafo de una caja que contiene 3 repuestos azules, 2 rojos y 3 verdes. Si X es el número de repuestos azules y Y es el número de repuestos rojos seleccionados, calcule

a) la función de probabilidad conjunta $f(x, y)$,

b) $P[(X, Y) \in A]$, donde A es la región $\{(x, y) \mid x + y \leq 1\}$.

***Solución*:** Los posibles pares de valores (x, y) son (0, 0), (0, 1), (1, 0), (1, 1), (0, 2) y (2, 0).

a) Ahora bien, $f(0, 1)$, por ejemplo, representa la probabilidad de seleccionar un repuesto rojo y uno verde. El número total de formas igualmente probables de seleccionar cualesquiera 2 repuestos de los 8 es $\binom{8}{2} = 28$. El número de formas de seleccionar 1 rojo de 2 repuestos rojos y 1 verde de 3 repuestos verdes es $\binom{2}{1}\binom{3}{1} = 6$. En consecuencia, $f(0, 1) = 6/28 = 3/14$. Cálculos similares dan las probabilidades para

los otros casos, los cuales se presentan en la tabla 3.1. Observe que las probabilidades suman 1. En el capítulo 5 se volverá evidente que la distribución de probabilidad conjunta de la tabla 3.1 se puede representar con la fórmula

$$f(x, y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}},$$

para $x = 0, 1, 2$; $y = 0, 1, 2$; y $0 \leq x + y \leq 2$.

b) La probabilidad de que (X,Y) caiga en la región A es

$$P[(X, Y) \in A] = P(X + Y \leq 1) = f(0,0) + f(0,1) + f(1,0)$$
$$= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}.$$

Tabla 3.1: Distribución de probabilidad conjunta para el ejemplo 3.14

		x			Totales
	$f(x, y)$	0	1	2	por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Cuando X y Y son variables aleatorias continuas, la **función de densidad conjunta** $f(x,y)$ es una superficie que yace sobre el plano xy, y $P[(X,Y) \in A]$, donde A es cualquier región en el plano xy, que es igual al volumen del cilindro recto limitado por la base A y la superficie.

Definición 3.9: La función $f(x,y)$ es una **función de densidad conjunta** de las variables aleatorias continuas X y Y si

1. $f(x, y) \geq 0$, para toda (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P[(X, Y) \in A] = \int\int_A f(x, y)\, dx\, dy$, para cualquier región A en el plano xy.

Ejemplo 3.15: Una empresa privada opera un local que da servicio a clientes que llegan en automóvil y otro que da servicio a clientes que llegan caminando. En un día elegido al azar, sean X y Y, respectivamente, las proporciones de tiempo que ambos locales están en servicio, y suponiendo que la función de densidad conjunta de estas variables aleatorias es

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique la condición 2 de la definición 3.9.

b) Calcule $P[(X, Y) \in A]$, donde $A = \{(x, y) \mid 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$.

***Solución*:** *a*) La integración de $f(x,y)$ sobre la totalidad de la región es

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,\ y)\ dx\ dy &= \int_0^1\int_0^1 \frac{2}{5}(2x+3y)\ dx\ dy\\ &= \int_0^1 \left(\frac{2x^2}{5}+\frac{6xy}{5}\right)\Big|_{x=0}^{x=1} dy\\ &= \int_0^1 \left(\frac{2}{5}+\frac{6y}{5}\right) dy = \left(\frac{2y}{5}+\frac{3y^2}{5}\right)\Big|_0^1 = \frac{2}{5}+\frac{3}{5}=1.\end{aligned}$$

b) Para calcular la probabilidad utilizamos

$$\begin{aligned}P[(X,\ Y)\in A] &= P\left(0<X<\frac{1}{2}, \frac{1}{4}<Y<\frac{1}{2}\right)\\ &= \int_{1/4}^{1/2}\int_0^{1/2}\frac{2}{5}(2x+3y)\ dx\ dy\\ &= \int_{1/4}^{1/2}\left(\frac{2x^2}{5}+\frac{6xy}{5}\right)\Big|_{x=0}^{x=1/2} dy = \int_{1/4}^{1/2}\left(\frac{1}{10}+\frac{3y}{5}\right) dy\\ &= \left(\frac{y}{10}+\frac{3y^2}{10}\right)\Big|_{1/4}^{1/2}\\ &= \frac{1}{10}\left[\left(\frac{1}{2}+\frac{3}{4}\right)-\left(\frac{1}{4}+\frac{3}{16}\right)\right] = \frac{13}{160}.\end{aligned}$$

Dada la distribución de probabilidad conjunta $f(x,y)$ de las variables aleatorias discretas X y Y, la distribución de probabilidad $g(x)$ sólo de X se obtiene sumando $f(x,y)$ sobre los valores de Y. De manera similar, la distribución de probabilidad $h(y)$ de sólo Y se obtiene sumando $f(x,y)$ sobre los valores de X. Definimos $g(x)$ y $h(y)$ como **distribuciones marginales** de X y Y, respectivamente. Cuando X y Y son variables aleatorias continuas, las sumatorias se reemplazan por integrales. Ahora podemos establecer la siguiente definición general.

Definición 3.10: Las **distribuciones marginales** sólo de X y sólo de Y son

$$g(x)=\sum_y f(x,\ y) \quad \text{y} \quad h(y)=\sum_x f(x,\ y)$$

para el caso discreto, y

$$g(x)=\int_{-\infty}^{\infty} f(x,\ y)\ dy \quad \text{y} \quad h(y)=\int_{-\infty}^{\infty} f(x,\ y)\,dx$$

para el caso continuo.

El término *marginal* se utiliza aquí porque, en el caso discreto, los valores de $g(x)$ y $h(y)$ son precisamente los totales marginales de las columnas y los renglones respectivos, cuando los valores de $f(x,y)$ se muestran en una tabla rectangular.

Ejemplo 3.16: Muestre que los totales de columnas y renglones de la tabla 3.1 dan las distribuciones marginales de sólo X y sólo Y.

***Solución*:** Para la variable aleatoria X vemos que

$$g(0) = f(0,0) + f(0,1) + f(0,2) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14},$$
$$g(1) = f(1,0) + f(1,1) + f(1,2) = \frac{9}{28} + \frac{3}{14} + 0 = \frac{15}{28},$$

y

$$g(2) = f(2,0) + f(2,1) + f(2,2) = \frac{3}{28} + 0 + 0 = \frac{3}{28},$$

que son precisamente los totales por columna de la tabla 3.1. De manera similar podemos mostrar que los valores de $h(y)$ están dados por los totales de los renglones. En forma tabular, estas distribuciones marginales se pueden escribir como sigue:

x	0	1	2
$g(x)$	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$

y	0	1	2
$h(y)$	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$

Ejemplo 3.17: Calcule $g(x)$ y $h(y)$ para la función de densidad conjunta del ejemplo 3.15.

***Solución*:** Por definición,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 \frac{2}{5}(2x + 3y)\, dy = \left(\frac{4xy}{5} + \frac{6y^2}{10}\right)\Bigg|_{y=0}^{y=1} = \frac{4x + 3}{5},$$

para $0 \le x \le 1$, y $g(x) = 0$ en otro caso. De manera similar,

$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^1 \frac{2}{5}(2x + 3y)\, dx = \frac{2(1 + 3y)}{5},$$

para $0 \le y \le 1$, y $h(y) = 0$ en otro caso.

El hecho de que las distribuciones marginales $g(x)$ y $h(y)$ sean en realidad las distribuciones de probabilidad de las variables individuales X y Y solas se puede verificar mostrando que se satisfacen las condiciones de la definición 3.4 o de la definición 3.6. Por ejemplo, en el caso continuo

$$\int_{-\infty}^{\infty} g(x)\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = 1,$$

y

$$P(a < X < b) = P(a < X < b, -\infty < Y < \infty)$$
$$= \int_a^b \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = \int_a^b g(x)\, dx.$$

En la sección 3.1 establecimos que el valor x de la variable aleatoria X representa un evento que es un subconjunto del espacio muestral. Si utilizamos la definición de probabilidad condicional que se estableció en el capítulo 2,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ siempre que } P(A) > 0,$$

donde A y B son ahora los eventos definidos por $X = x$ y $Y = y$, respectivamente, entonces,

$$P(Y = y \mid X = x) = \frac{P(X = x,\ Y = y)}{P(X = x)} = \frac{f(x,\ y)}{g(x)},\ \text{siempre que}\ g(x) > 0,$$

donde X y Y son variables aleatorias discretas.

No es difícil mostrar que la función $f(x, y)/g(x)$, que es estrictamente una función de y con x fija, satisface todas las condiciones de una distribución de probabilidad. Esto también es cierto cuando $f(x, y)$ y $g(x)$ son la densidad conjunta y la distribución marginal, respectivamente, de variables aleatorias continuas. Como resultado, para poder calcular probabilidades condicionales de manera eficaz es sumamente importante que utilicemos el tipo especial de distribución de la forma $f(x, y)/g(x)$. Este tipo de distribución se llama **distribución de probabilidad condicional** y se define formalmente como sigue:

Definición 3.11: Sean X y Y dos variables aleatorias, discretas o continuas. La **distribución condicional** de la variable aleatoria Y, dado que $X = x$, es

$$f(y|x) = \frac{f(x,\ y)}{g(x)},\ \text{siempre que}\ g(x) > 0.$$

De manera similar, la distribución condicional de la variable aleatoria X, dado que $Y = y$, es

$$f(x|y) = \frac{f(x,\ y)}{h(y)},\ \text{siempre que}\ h(y) > 0.$$

Si deseamos encontrar la probabilidad de que la variable aleatoria discreta X caiga entre a y b cuando sabemos que la variable discreta $Y = y$, evaluamos

$$P(a < X < b \mid Y = y) = \sum_{a<x<b} f(x|y),$$

donde la sumatoria se extiende a todos los valores de X entre a y b. Cuando X y Y son continuas, evaluamos

$$P(a < X < b \mid Y = y) = \int_a^b f(x|y)\ dx.$$

Ejemplo 3.18: Remítase al ejemplo 3.14, calcule la distribución condicional de X, dado que $Y = 1$, y utilice el resultado para determinar $P(X = 0 \mid Y = 1)$.

***Solución*:** Necesitamos encontrar $f(x \mid y)$, donde $y = 1$. Primero calculamos que

$$h(1) = \sum_{x=0}^{2} f(x,\ 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}.$$

Ahora calculamos,

$$f(x|1) = \frac{f(x,\ 1)}{h(1)} = \left(\frac{7}{3}\right) f(x,\ 1), \quad x = 0, 1, 2.$$

Por lo tanto,

$$f(0|1) = \left(\frac{7}{3}\right) f(0, 1) = \left(\frac{7}{3}\right)\left(\frac{3}{14}\right) = \frac{1}{2},\ f(1|1) = \left(\frac{7}{3}\right) f(1, 1) = \left(\frac{7}{3}\right)\left(\frac{3}{14}\right) = \frac{1}{2},$$

$$f(2|1) = \left(\frac{7}{3}\right) f(2, 1) = \left(\frac{7}{3}\right)(0) = 0,$$

y la distribución condicional de X, dado que $Y = 1$, es

x	0	1	2
$f(x\|1)$	$\frac{1}{2}$	$\frac{1}{2}$	0

Finalmente,

$$P(X = 0 \mid Y = 1) = f(0|1) = \frac{1}{2}.$$

Por lo tanto, si se sabe que 1 de los 2 repuestos seleccionados es rojo, tenemos una probabilidad igual a $1/2$ de que el otro repuesto no sea azul. ■

Ejemplo 3.19: La densidad conjunta para las variables aleatorias (X,Y), donde X es el cambio unitario de temperatura y Y es la proporción de desplazamiento espectral que produce cierta partícula atómica es

$$f(x, y) = \begin{cases} 10xy^2, & 0 < x < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule las densidades marginales $g(x)$, $h(y)$ y la densidad condicional $f(y\,|\,x)$.

b) Calcule la probabilidad de que el espectro se desplace más de la mitad de las observaciones totales, dado que la temperatura se incremente en 0.25 unidades.

***Solución*:** *a*) Por definición,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{x}^{1} 10xy^2\, dy$$

$$= \frac{10}{3} xy^3 \Big|_{y=x}^{y=1} = \frac{10}{3} x(1 - x^3),\ 0 < x < 1,$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{0}^{y} 10xy^2\, dx = 5x^2y^2 \Big|_{x=0}^{x=y} = 5y^4,\ 0 < y < 1.$$

Entonces,

$$f(y|x) = \frac{f(x, y)}{g(x)} = \frac{10xy^2}{\frac{10}{3}x(1 - x^3)} = \frac{3y^2}{1 - x^3},\ 0 < x < y < 1.$$

b) Por lo tanto,

$$P\left(Y > \frac{1}{2} \,\middle|\, X = 0.25\right) = \int_{1/2}^{1} f(y \mid x = 0.25)\, dy = \int_{1/2}^{1} \frac{3y^2}{1 - 0.25^3}\, dy = \frac{8}{9}.$$ ■

Ejemplo 3.20: Dada la función de densidad conjunta

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule $g(x)$, $h(y)$, $f(x\,|\,y)$ y evalúe $P(\frac{1}{4} < X < \frac{1}{2} \mid Y = \frac{1}{3})$.

***Solución*:** Por definición de la densidad marginal, para $0 < x < 2$,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 \frac{x(1+3y^2)}{4} dy$$
$$= \left(\frac{xy}{4} + \frac{xy^3}{4} \right) \Bigg|_{y=0}^{y=1} = \frac{x}{2},$$

y para $0 < y < 1$,

$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^2 \frac{x(1+3y^2)}{4} dx$$
$$= \left(\frac{x^2}{8} + \frac{3x^2y^2}{8} \right) \Bigg|_{x=0}^{x=2} = \frac{1+3y^2}{2}.$$

Por lo tanto, usando la definición de la densidad condicional para $0 < x < 2$,

$$f(x|y) = \frac{f(x,y)}{h(y)} = \frac{x(1+3y^2)/4}{(1+3y^2)/2} = \frac{x}{2},$$

y

$$P\left(\frac{1}{4} < X < \frac{1}{2} \;\middle|\; Y = \frac{1}{3} \right) = \int_{1/4}^{1/2} \frac{x}{2} dx = \frac{3}{64}.$$

Independencia estadística

Si $f(x|y)$ no depende de y, como ocurre en el ejemplo 3.20, entonces $f(x|y) = g(x)$ y $f(x,y) = g(x)h(y)$. La prueba se realiza sustituyendo

$$f(x, y) = f(x|y)h(y)$$

en la distribución marginal de X. Es decir,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{\infty} f(x|y)h(y)\, dy.$$

Si $f(x|y)$ no depende de y, podemos escribir

$$g(x) = f(x|y) \int_{-\infty}^{\infty} h(y)\, dy.$$

Entonces,

$$\int_{-\infty}^{\infty} h(y)\, dy = 1,$$

ya que $h(y)$ es la función de densidad de probabilidad de Y. Por lo tanto,

$$g(x) = f(x|y) \text{ y entonces } f(x, y) = g(x)h(y).$$

Debería tener sentido para el lector que si $f(x|y)$ no depende de y, entonces, por supuesto, el resultado de la variable aleatoria Y no repercute en el resultado de la variable aleatoria X. En otras palabras, decimos que X y Y son variables aleatorias independientes. Ofrecemos ahora la siguiente definición formal de independencia estadística.

Definición 3.12: Sean X y Y dos variables aleatorias, discretas o continuas, con distribución de probabilidad conjunta $f(x,y)$ y distribuciones marginales $g(x)$ y $h(y)$, respectivamente. Se dice que las variables aleatorias X y Y son **estadísticamente independientes** si y sólo si

$$f(x, y) = g(x)h(y)$$

para toda (x,y) dentro de sus rangos.

Las variables aleatorias continuas del ejemplo 3.20 son estadísticamente independientes, pues el producto de las dos distribuciones marginales da la función de densidad conjunta. Sin embargo, es evidente que ése no es el caso de las variables continuas del ejemplo 3.19. La comprobación de la independencia estadística de variables aleatorias discretas requiere una investigación más profunda, ya que es posible que el producto de las distribuciones marginales sea igual a la distribución de probabilidad conjunta para algunas, aunque no para todas, las combinaciones de (x,y). Si puede encontrar algún punto (x,y) para el que $f(x,y)$ se define de manera que $f(x,y) \neq g(x)h(y)$, las variables discretas X y Y no son estadísticamente independientes.

Ejemplo 3.21: Demuestre que las variables aleatorias del ejemplo 3.14 no son estadísticamente independientes.

***Prueba*:** Consideremos el punto (0,1). A partir de la tabla 3.1, encontramos que las tres probabilidades $f(0,1)$, $g(0)$ y $h(1)$ son

$$f(0, 1) = \frac{3}{14},$$

$$g(0) = \sum_{y=0}^{2} f(0, y) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14},$$

$$h(1) = \sum_{x=0}^{2} f(x, 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}.$$

Claramente,

$$f(0, 1) \neq g(0)h(1),$$

por lo tanto, X y Y no son estadísticamente independientes. ∎

Todas las definiciones anteriores respecto a dos variables aleatorias se pueden generalizar al caso de n variables aleatorias. Sea $f(x_1, x_2, \ldots, x_n)$ la función de probabilidad conjunta de las variables aleatorias $X_1, X_2, \ldots, X_n$. La distribución marginal de X_1, por ejemplo, es

$$g(x_1) = \sum_{x_2} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n)$$

para el caso discreto, y

$$g(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_2\, dx_3 \cdots dx_n$$

para el caso continuo. Ahora podemos obtener **distribuciones marginales conjuntas** como $g(x_1, x_2)$, donde

$$g(x_1, x_2) = \begin{cases} \sum_{x_3} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n) & \text{(caso discreto),} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_3\, dx_4 \cdots dx_n & \text{(caso continuo).} \end{cases}$$

Podríamos considerar numerosas distribuciones condicionales. Por ejemplo, la **distribución condicional conjunta** de X_1, X_2 y X_3, dado que $X_4 = x_4$, $X_5 = x_5, \ldots, X_n = x_n$, se escribe como

$$f(x_1, x_2, x_3 \mid x_4, x_5, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{g(x_4, x_5, \ldots, x_n)},$$

donde $g(x_4, x_5, \ldots, x_n)$ es la distribución marginal conjunta de las variables aleatorias X_4, $X_5, \ldots, X_n$.

Una generalización de la definición 3.12 nos lleva a la siguiente definición para la independencia estadística mutua de las variables $X_1, X_2, \ldots, X_n$.

Definición 3.13: Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias, discretas o continuas, con distribución de probabilidad conjunta $f(x_1, x_2, \ldots, x_n)$ y distribuciones marginales $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectivamente. Se dice que las variables aleatorias $X_1, X_2, \ldots, X_n$ son recíproca y **estadísticamente independientes** si y sólo si

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$$

para toda $(x_1, x_2, \ldots, x_n)$ dentro de sus rangos.

Ejemplo 3.22: Suponga que el tiempo de vida en anaquel de cierto producto comestible perecedero empacado en cajas de cartón, en años, es una variable aleatoria cuya función de densidad de probabilidad está dada por

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Represente los tiempos de vida en anaquel para tres de estas cajas seleccionadas de forma independiente con X_1, X_2 y X_3 y calcule $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$.

***Solución*:** Como las cajas se seleccionaron de forma independiente, suponemos que las variables aleatorias X_1, X_2 y X_3 son estadísticamente independientes y que tienen la siguiente densidad de probabilidad conjunta:

$$f(x_1, x_2, x_3) = f(x_1) f(x_2) f(x_3) = e^{-x_1} e^{-x_2} e^{-x_3} = e^{-x_1 - x_2 - x_3},$$

para $x_1 > 0$, $x_2 > 0$, $x_3 > 0$, y $f(x_1, x_2, x_3) = 0$ en cualquier otro caso. En consecuencia,

$$\begin{aligned} P(X_1 < 2, 1 < X_2 < 3, X_3 > 2) &= \int_2^{\infty} \int_1^3 \int_0^2 e^{-x_1 - x_2 - x_3}\, dx_1\, dx_2\, dx_3 \\ &= (1 - e^{-2})(e^{-1} - e^{-3})e^{-2} = 0.0372. \end{aligned}$$

¿Por qué son importantes las características de las distribuciones de probabilidad y de dónde provienen?

Es importante que este texto ofrezca al lector una transición hacia los siguientes tres capítulos. En los ejemplos y los ejercicios presentamos casos de situaciones prácticas de ingeniería y ciencias, en los cuales las distribuciones de probabilidad y sus propiedades se utilizan para resolver problemas importantes. Estas distribuciones de probabilidad, ya sean discretas o continuas, se presentaron mediante frases como "se sabe que", "suponga que" o incluso, en ciertos casos, "la evidencia histórica sugiere que". Se trata de situaciones en las que la naturaleza de la distribución, e incluso una estimación óptima de la estructura de la probabilidad, se pueden determinar utilizando datos históricos, datos tomados de estudios a largo plazo o incluso de grandes cantidades de datos planeados. El lector debería tener presente lo expuesto en el capítulo 1 respecto al uso de histogramas y, por consiguiente, recordar cómo se estiman las distribuciones de frecuencias a partir de los histogramas. Sin embargo, no todas las funciones de probabilidad y de densidad de probabilidad se derivan de cantidades grandes de datos históricos. Hay un gran número de situaciones en las que la naturaleza del escenario científico sugiere un tipo de distribución. De hecho, varias de ellas se reflejan en los ejercicios del capítulo 2 y en este capítulo. Cuando observaciones repetidas independientes son binarias por naturaleza (es decir, defectuoso o no, funciona o no, alérgico o no) con un valor de 0 o 1, la distribución que cubre esta situación se llama **distribución binomial**. La función de probabilidad de esta distribución se explicará y se demostrará en el capítulo 5. El ejercicio 3.34 de la sección 3.3 y el ejercicio de repaso 3.80 constituyen ejemplos de este tipo de distribución, y hay otros que el lector también debería reconocer. El escenario de una distribución continua del tiempo de operación antes de cualquier falla, como en el ejercicio de repaso 3.69 o en el ejercicio 3.27 de la página 93, a menudo sugiere una clase de distribución denominada **distribución exponencial**. Tales tipos de ejemplos son tan sólo dos de la gran cantidad de las llamadas distribuciones estándar que se utilizan ampliamente en situaciones del mundo real porque el escenario científico que da lugar a cada uno de ellos es reconocible y a menudo se presenta en la práctica. Los capítulos 5 y 6 abarcan muchos de estos tipos de ejemplos, junto con alguna teoría inherente respecto de su uso.

La segunda parte de esta transición al material de los capítulos siguientes tiene que ver con el concepto de **parámetros de la población** o **parámetros de distribución**. Recuerde que en el capítulo 1 analizamos la necesidad de utilizar datos para ofrecer información sobre dichos parámetros. Profundizamos en el estudio de las nociones de **media** y de **varianza**, y proporcionamos ideas sobre esos conceptos en el contexto de una población. De hecho, es fácil calcular la media y la varianza de la población a partir de la función de probabilidad para el caso discreto, o de la función de densidad de probabilidad para el caso continuo. Tales parámetros y su importancia en la solución de muchas clases de problemas de la vida real nos proporcionarán gran parte del material de los capítulos 8 a 17.

Ejercicios

3.37 Determine los valores de c, tales que las siguientes funciones representen distribuciones de probabilidad conjunta de las variables aleatorias X y Y:

a) $f(x, y) = cxy$, para $x = 1, 2, 3$; $y = 1, 2, 3$;

b) $f(x, y) = c\,|x - y|$, para $x = -2, 0, 2$; $y = -2, 3$.

3.38 Si la distribución de probabilidad conjunta de X y Y está dada por

$$f(x, y) = \frac{x + y}{30}, \quad \text{para} \quad x = 0, 1, 2, 3; \; y = 0, 1, 2,$$

calcule

a) $P(X \leq 2, Y = 1)$;
b) $P(X > 2, Y \leq 1)$;
c) $P(X > Y)$;
d) $P(X + Y = 4)$.

3.39 De un saco de frutas que contiene 3 naranjas, 2 manzanas y 3 plátanos se selecciona una muestra aleatoria de 4 frutas. Si X es el número de naranjas y Y el de manzanas en la muestra, calcule

a) la distribución de probabilidad conjunta de X y Y;
b) $P[(X, Y) \in A]$, donde A es la región dada por $\{(x, y) \mid x + y \leq 2\}$.

3.40 Un restaurante de comida rápida opera tanto en un local que da servicio en el automóvil, como en un local que atiende a los clientes que llegan caminando. En un día elegido al azar, represente las proporciones de tiempo que el primero y el segundo local están en servicio con X y Y, respectivamente, y suponga que la función de densidad conjunta de estas variables aleatorias es

$$f(x, y) = \begin{cases} \frac{2}{3}(x + 2y), & 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule la densidad marginal de X.
b) Calcule la densidad marginal de Y.
c) Calcule la probabilidad de que el local que da servicio a los clientes que llegan en automóvil esté lleno menos de la mitad del tiempo.

3.41 Una empresa dulcera distribuye cajas de chocolates con un surtido de cremas, chiclosos y envinados. Suponga que cada caja pesa 1 kilogramo, pero que los pesos individuales de cremas, chiclosos y envinados varían de una a otra cajas. Para una caja seleccionada al azar, represente los pesos de las cremas y los chiclosos con X y Y, respectivamente, y suponga que la función de densidad conjunta de estas variables es

$$f(x, y) = \begin{cases} 24xy, & 0 \leq x \leq 1,\ 0 \leq y \leq 1,\ x + y \leq 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la probabilidad de que en una caja dada los envinados representen más de la mitad del peso.
b) Calcule la densidad marginal para el peso de las cremas.
c) Calcule la probabilidad de que el peso de los chiclosos en una caja sea menor que $1/8$ de kilogramo, si se sabe que las cremas constituyen $3/4$ partes del peso.

3.42 Sean X y Y la duración de la vida, en años, de dos componentes en un sistema electrónico. Si la función de densidad conjunta de estas variables es

$$f(x, y) = \begin{cases} e^{-(x+y)}, & x > 0,\ y > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule $P(0 < X < 1 \mid Y = 2)$.

3.43 Sea X el tiempo de reacción, en segundos, ante cierto estímulo, y Y la temperatura (en °F) a la cual inicia cierta reacción. Suponga que dos variales aleatorias, X y Y, tienen la densidad conjunta

$$f(x, y) = \begin{cases} 4xy, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule

a) $P(0 \leq X \leq \frac{1}{2} \text{ y } \frac{1}{4} \leq Y \leq \frac{1}{2})$;
b) $P(X < Y)$.

3.44 Se supone que cada rueda trasera de un avión experimental se llena a una presión de 40 libras por pulgada cuadrada (psi). Sea X la presión real del aire para la rueda derecha y Y la presión real del aire de la rueda izquierda. Suponga que X y Y son variables aleatorias con la siguiente función de densidad conjunta:

$$f(x, y) = \begin{cases} k(x^2 + y^2), & 30 \leq x < 50,\ 30 \leq y < 50, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule k.
b) Calcule $P(30 \leq X \leq 40 \text{ y } 40 \leq Y < 50)$.
c) Calcule la probabilidad de que ambas ruedas no contengan la suficiente cantidad de aire.

3.45 Sea X el diámetro de un cable eléctrico blindado y Y el diámetro del molde cerámico que hace el cable. Tanto X como Y tienen una escala tal que están entre 0 y 1. Suponga que X y Y tienen la siguiente densidad conjunta:

$$f(x, y) = \begin{cases} \frac{1}{y}, & 0 < x < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule $P(X + Y > 1/2)$.

3.46 Remítase al ejercicio 3.38, calcule

a) la distribución marginal de X;
b) la distribución marginal de Y.

3.47 Al principio de cualquier día la cantidad de queroseno que contiene un tanque, en miles de litros, es una cantidad aleatoria Y, de la que durante el día se vende una cantidad aleatoria X. Suponga que el tanque no se reabastece durante el día, de manera que $x \leq y$, e imagine también que la función de densidad conjunta de estas variables es

$$f(x, y) = \begin{cases} 2, & 0 < x \leq y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine si X y Y son independientes.
b) Calcule $P(1/4 < X < 1/2 \mid Y = 3/4)$.

3.48 Remítase al ejercicio 3.39 y calcule
a) $f(y \mid 2)$ para todos los valores de y;
b) $P(Y = 0 \mid X = 2)$.

3.49 Sea X el número de veces que fallará cierta máquina de control numérico: 1, 2 o 3 veces en un día dado. Y si Y denota el número de veces que se llama a un técnico para una emergencia, su distribución de probabilidad conjunta estará dada como

$f(x, y)$		x = 1	x = 2	x = 3
y	1	0.05	0.05	0.10
	3	0.05	0.10	0.35
	5	0.00	0.20	0.10

a) Evalúe la distribución marginal de X.
b) Evalúe la distribución marginal de Y.
c) Calcule $P(Y = 3 \mid X = 2)$.

3.50 Suponga que X y Y tienen la siguiente distribución de probabilidad conjunta:

$f(x, y)$		x = 2	x = 4
y	1	0.10	0.15
	3	0.20	0.30
	5	0.10	0.15

a) Calcule la distribución marginal de X.
b) Calcule la distribución marginal de Y.

3.51 De las 12 cartas mayores (jotas, reinas y reyes) de una baraja ordinaria de 52 cartas se sacan tres cartas sin reemplazo. Sea X el número de reyes que se seleccionan y Y el número de jotas. Calcule
a) la distribución de probabilidad conjunta de X y Y;
b) $P[(X,Y) \in A]$, donde A es la región dada por $\{(x,y) \mid x + y \geq 2\}$.

3.52 Una moneda se lanza dos veces. Sea Z el número de caras en el primer lanzamiento y W el número total de caras en los 2 lanzamientos. Si la moneda no está balanceada y una cara tiene una probabilidad de ocurrencia de 40%, calcule
a) la distribución de probabilidad conjunta de W y Z;
b) la distribución marginal de W;
c) la distribución marginal de Z;
d) la probabilidad de que ocurra al menos 1 cara.

3.53 Dada la función de densidad conjunta

$$f(x, y) = \begin{cases} \frac{6-x-y}{8}, & 0 < x < 2,\ 2 < y < 4, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule $P(1 < Y < 3 \mid X = 1)$.

3.54 Determine si las dos variables aleatorias del ejercicio 3.49 son dependientes o independientes.

3.55 Determine si las dos variables aleatorias del ejercicio 3.50 son dependientes o independientes.

3.56 La función de densidad conjunta de las variables aleatorias X y Y es

$$f(x, y) = \begin{cases} 6x, & 0 < x < 1,\ 0 < y < 1 - x, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Demuestre que X y Y no son independientes.
b) Calcule $P(X > 0.3 \mid Y = 0.5)$.

3.57 Si X, Y y Z tienen la siguiente función de densidad de probabilidad conjunta:

$$f(x, y, z) = \begin{cases} kxy^2z, & 0 < x, y < 1,\ 0 < z < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule k.
b) Calcule $P(X < \frac{1}{4}, Y > \frac{1}{2}, 1 < Z < 2)$.

3.58 Determine si las dos variables aleatorias del ejercicio 3.43 son dependientes o independientes.

3.59 Determine si las dos variables aleatorias del ejercicio 3.44 son dependientes o independientes.

3.60 La función de densidad de probabilidad conjunta de las variables aleatorias X, Y y Z es

$$f(x, y, z) = \begin{cases} \frac{4xyz^2}{9}, & 0 < x, y < 1,\ 0 < z < 3, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule
a) la función de densidad marginal conjunta de Y y Z;
b) la densidad marginal de Y;
c) $P(\frac{1}{4} < X < \frac{1}{2},\ Y > \frac{1}{3},\ 1 < Z < 2)$;
d) $P(0 < X < \frac{1}{2} \mid Y = \frac{1}{4},\ Z = 2)$.

Ejercicios de repaso

3.61 Una empresa tabacalera produce mezclas de tabaco. Cada mezcla contiene diversas proporciones de tabaco turco, tabaco de la región y otros. Las proporciones de tabaco turco y de la región en una mezcla son variables aleatorias con una función de densidad conjunta (X = turco y Y = de la región)

$$f(x, y) = \begin{cases} 24xy, & 0 \le x, y \le 1,\ x + y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule la probabilidad de que en determinada caja el tabaco turco represente más de la mitad de la mezcla.
b) Calcule la función de densidad marginal para la proporción del tabaco de la región.
c) Calcule la probabilidad de que la proporción de tabaco turco sea menor que $1/8$, si se sabe que la mezcla contiene $3/4$ de tabaco de la región.

3.62 Una empresa de seguros ofrece a sus asegurados varias opciones diferentes de pago de la prima. Para un asegurado seleccionado al azar, sea X el número de meses entre pagos sucesivos. La función de distribución acumulada de X es

$$F(x) = \begin{cases} 0, & \text{si } x < 1, \\ 0.4, & \text{si } 1 \le x < 3, \\ 0.6, & \text{si } 3 \le x < 5, \\ 0.8, & \text{si } 5 \le x < 7, \\ 1.0, & \text{si } x \ge 7. \end{cases}$$

a) ¿Cuál es la función de masa de probabilidad de X?
b) Calcule $P(4 < X \le 7)$.

3.63 Dos componentes electrónicos de un sistema de proyectiles funcionan en conjunto para el éxito de todo el sistema. Sean X y Y la vida en horas de los dos componentes. La densidad conjunta de X y Y es

$$f(x, y) = \begin{cases} ye^{-y(1+x)}, & x, y \ge 0, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine las funciones de densidad marginal para ambas variables aleatorias.
b) ¿Cuál es la probabilidad de que ambos componentes duren más de dos horas?

3.64 Una instalación de servicio opera con dos líneas telefónicas. En un día elegido al azar, sea X la proporción de tiempo que la primera línea está en uso, mientras que Y es la proporción de tiempo en que la segunda línea está en uso. Suponga que la función de densidad de probabilidad conjunta para (X,Y) es

$$f(x, y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 \le x, y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule la probabilidad de que ninguna línea esté ocupada más de la mitad del tiempo.
b) Calcule la probabilidad de que la primera línea esté ocupada más del 75% del tiempo.

3.65 Sea el número de llamadas telefónicas que recibe un conmutador durante un intervalo de 5 minutos una variable aleatoria X con la siguiente función de probabilidad:

$$f(x) = \frac{e^{-2}2^x}{x!}, \quad \text{para} \quad x = 0, 1, 2, \ldots.$$

a) Determine la probabilidad de que X sea igual a 0, 1, 2, 3, 4, 5 y 6.
b) Grafique la función de masa de probabilidad para estos valores de x.
c) Determine la función de distribución acumulada para estos valores de X.

3.66 Considere las variables aleatorias X y Y con la siguiente función de densidad conjunta

$$f(x, y) = \begin{cases} x + y, & 0 \le x, y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Calcule las distribuciones marginales de X y Y.
b) Calcule $P(X > 0.5, Y > 0.5)$.

3.67 En un proceso industrial se elaboran artículos que se pueden clasificar como defectuosos o no defectuosos. La probabilidad de que un artículo esté defectuoso es 0.1. Se realiza un experimento en el que 5 artículos del proceso se eligen al azar. Sea la variable aleatoria X el número de artículos defectuosos en esta muestra de 5. ¿Cuál es la función de masa de probabilidad de X?

3.68 Considere la siguiente función de densidad de probabilidad conjunta de las variables aleatorias X y Y:

$$f(x, y) = \begin{cases} \frac{3x - y}{9}, & 1 < x < 3,\ 1 < y < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule las funciones de densidad marginal de X y Y.
b) ¿X y Y son independientes?
c) Calcule $P(X > 2)$.

3.69 La duración en horas de un componente eléctrico es una variable aleatoria con la siguiente función de distribución acumulada:

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{50}}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine su función de densidad de probabilidad.
b) Determine la probabilidad de que la vida útil de tal componente rebase las 70 horas.

3.70 En una fábrica específica de pantalones un grupo de 10 trabajadores los inspecciona tomando aleatoriamente algunos de la línea de producción. A cada inspector se le asigna un número del 1 al 10. Un comprador selecciona un pantalón para adquirirlo. Sea la variable aleatoria X el número del inspector.
a) Determine una función de masa de probabilidad razonable para X.
b) Grafique la función de distribución acumulada para X.

3.71 La vida en anaquel de un producto es una variable aleatoria que se relaciona con la aceptación por parte del consumidor. Resulta que la vida en anaquel Y, en días, de cierta clase de artículo de panadería tiene la siguiente función de densidad:

$$f(y) = \begin{cases} \frac{1}{2}e^{-y/2}, & 0 \le y < \infty, \\ 0, & \text{en otro caso.} \end{cases}$$

¿Qué fracción de las rebanadas de este producto que hoy están en exhibición se espera que se vendan en 3 días a partir de hoy?

3.72 El congestionamiento de pasajeros es un problema de servicio en los aeropuertos, en los cuales se instalan trenes para reducir la congestión. Cuando se usa el tren, el tiempo X, en minutos, que toma viajar desde la terminal principal hasta una explanada específica tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Demuestre que la función de densidad de probabilidad anterior es válida.
b) Calcule la probabilidad de que el tiempo que le toma a un pasajero viajar desde la terminal principal hasta la explanada no exceda los 7 minutos.

3.73 Las impurezas en el lote del producto final de un proceso químico a menudo reflejan un grave problema. A partir de una cantidad considerable de datos recabados en la planta se sabe que la proporción Y de impurezas en un lote tiene una función de densidad dada por

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que la función de densidad anterior sea válida.
b) Se considera que un lote no es vendible y, por consiguiente, no es aceptable si el porcentaje de impurezas es superior a 60%. Con la calidad del proceso actual, ¿cuál es el porcentaje de lotes que no son aceptables?

3.74 El tiempo Z, en minutos, entre llamadas a un sistema de alimentación eléctrica tiene la siguiente función de densidad de probabilidad:

$$f(z) = \begin{cases} \frac{1}{10}e^{-z/10}, & 0 < z < \infty, \\ 0, & \text{en otro caso.} \end{cases}$$

a) ¿Cuál es la probabilidad de que no haya llamadas en un lapso de 20 minutos?
b) ¿Cuál es la probabilidad de que la primera llamada entre en los primeros 10 minutos después de abrir?

3.75 Un sistema químico que surge de una reacción química tiene dos componentes importantes, entre otros, en una mezcla. La distribución conjunta que describe las proporciones X_1 y X_2 de estos dos componentes está dada por

$$f(x_1, x_2) = \begin{cases} 2, & 0 < x_1 < x_2 < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine la distribución marginal de X_1.
b) Determine la distribución marginal de X_2.
c) ¿Cuál es la probabilidad de que las proporciones del componente generen los resultados $X_1 < 0.2$ y $X_2 > 0.5$?
d) Determine la distribución condicional $f_{X_1|X_2}(x_1 \mid x_2)$.

3.76 Considere la situación del ejercicio de repaso 3.75; pero suponga que la distribución conjunta de las dos proporciones está dada por

$$f(x_1, x_2) = \begin{cases} 6x_2, & 0 < x_2 < x_1 < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine la distribución marginal $f_{X_1}(x_1)$ de la proporción X_1 y verifique que sea una función de densidad válida.
b) ¿Cuál es la probabilidad de que la proporción X_2 sea menor que 0.5, dado que X_1 es 0.7?

3.77 Considere las variables aleatorias X y Y que representan el número de vehículos que llegan a dos esquinas de calles separadas durante cierto periodo de 2 minutos. Estas esquinas de las calles están bastante cerca una de la otra, así que es importante que los ingenieros de tráfico se ocupen de ellas de manera conjunta si fuera necesario. Se sabe que la distribución conjunta de X y Y es

$$f(x, y) = \frac{9}{16} \cdot \frac{1}{4^{(x+y)}},$$

para $x = 0, 1, 2, \ldots$, y para $y = 0, 1, 2, \ldots$
a) ¿Son independientes las dos variables aleatorias X y Y? Explique su respuesta.

b) ¿Cuál es la probabilidad de que, durante el periodo en cuestión, lleguen menos de 4 vehículos a las dos esquinas?

3.78 El comportamiento de series de componentes desempeña un papel importante en problemas de confiabilidad científicos y de ingeniería. Ciertamente la confiabilidad de todo el sistema no es mejor que la del componente más débil de las series. En un sistema de series los componentes funcionan de manera independiente unos de otros. En un sistema particular de tres componentes, la probabilidad de cumplir con la especificación para los componentes 1, 2 y 3, respectivamente, son 0.95, 0.99 y 0.92. ¿Cuál es la probabilidad de que todo el sistema funcione?

3.79 Otro tipo de sistema que se utiliza en trabajos de ingeniería es un grupo de componentes en paralelo o un sistema paralelo. En este enfoque más conservador la probabilidad de que el sistema funcione es mayor que la probabilidad de que cualquier componente funcione. El sistema fallará sólo cuando falle todo el sistema. Considere una situación en la que hay 4 componentes independientes en un sistema paralelo, en la que la probabilidad de operación está dada por

Componente 1: 0.95; Componente 2: 0.94;
Componente 3: 0.90; Componente 4: 0.97.

¿Cuál es la probabilidad de que no falle el sistema?

3.80 Considere un sistema de componentes en que hay cinco componentes independientes, cada uno de los cuales tiene una probabilidad de operación de 0.92. De hecho, el sistema tiene una redundancia preventiva diseñada para que no falle mientras 3 de sus 5 componentes estén en funcionamiento. ¿Cuál es la probabilidad de que funcione todo el sistema?

3.81 **Proyecto de grupo:** Observe el color de los zapatos de los estudiantes en 5 periodos de clases. Suponga que las categorías de color son rojo, blanco, negro, café y otro. Construya una tabla de frecuencias para cada color.

a) Estime e interprete el significado de la distribución de probabilidad.

b) ¿Cuál es la probabilidad estimada de que en el siguiente periodo de clases un estudiante elegido al azar use un par de zapatos rojos o blancos?

3.5 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

En los siguientes capítulos será evidente que las distribuciones de probabilidad representan la estructura mediante la cual las probabilidades que se calculan ayudan a evaluar y a comprender un proceso. Por ejemplo, en el ejercicio de repaso 3.65 la distribución de probabilidad que cuantifica la probabilidad de que haya una carga excesiva durante ciertos periodos podría ser muy útil en la planeación de cualquier cambio en el sistema. El ejercicio de repaso 3.69 describe un escenario donde se estudia el periodo de vida útil de un componente electrónico. Conocer la estructura de la probabilidad para el componente contribuirá de manera significativa al entendimiento de la confiabilidad de un sistema mayor del cual éste forme parte. Además, comprender la naturaleza general de las distribuciones de probabilidad reforzará el conocimiento del concepto **valor-*P***, que se estudió brevemente en el capítulo 1 y que desempeñará un papel destacado al inicio del capítulo 10 y en lo que resta del texto.

Los capítulos 4, 5 y 6 dependen mucho del material cubierto en este capítulo. En el capítulo 4 estudiaremos el significado de **parámetros** importantes en las distribuciones de probabilidad. Tales parametros cuantifican las nociones de **tendencia central** y **variabilidad** en un sistema. De hecho, el conocimiento de tales cantidades, al margen de la distribución completa, puede ofrecer información sobre la naturaleza del sistema. En los capítulos 5 y 6 se examinarán escenarios de ingeniería, biológicos y de ciencia en general que identifican tipos de distribuciones especiales. Por ejemplo, la estructura de la función de probabilidad en el ejercicio de repaso 3.65 se identificará fácilmente bajo ciertas suposiciones que se estudiarán en el capítulo 5. Lo mismo ocurre en el contexto

del ejercicio de repaso 3.69, que es un caso especial de problema sobre **tiempo de operación antes de la falla**, cuya función de densidad de probabilidad se estudiará en el capítulo 6.

En lo que concierne a los riesgos potenciales de utilizar el material de este capítulo, la advertencia para el lector sería no interpretar el material más allá de lo que sea evidente. La naturaleza general de la distribución de probabilidad para un fenómeno científico determinado no es obvia a partir de lo que se estudió aquí. La finalidad de este capítulo es que los lectores aprendan a manipular una distribución de probabilidad, no que aprendan a identificar un tipo específico. Los capítulos 5 y 6 avanzan un largo trecho hacia la identificación de acuerdo con la naturaleza general del sistema científico.

Capítulo 4

Esperanza matemática

4.1 Media de una variable aleatoria

En el capítulo 1 estudiamos la media muestral, que es la media aritmética de los datos. Ahora considere la siguiente situación: si dos monedas se lanzan 16 veces y X es el número de caras que resultan en cada lanzamiento, entonces los valores de X pueden ser 0, 1 y 2. Suponga que los resultados del experimento son: cero caras, una cara y dos caras, un total de 4, 7 y 5 veces, respectivamente. El número promedio de caras por lanzamiento de las dos monedas es, entonces,

$$\frac{(0)(4) + (1)(7) + (2)(5)}{16} = 1.06.$$

Éste es un valor promedio de los datos, aunque no es un resultado posible de {0, 1, 2}. Por lo tanto, un promedio no es necesariamente un resultado posible del experimento. Por ejemplo, es probable que el ingreso mensual promedio de un vendedor no sea igual a alguno de sus cheques de pago mensuales.

Reestructuremos ahora nuestro cálculo del número promedio de caras para tener la siguiente forma equivalente:

$$(0)\left(\frac{4}{16}\right) + (1)\left(\frac{7}{16}\right) + (2)\left(\frac{5}{16}\right) = 1.06.$$

Los números 4/16, 7/16 y 5/16 son las fracciones de los lanzamientos totales que dan como resultado 0, 1 y 2 caras, respectivamente. Tales fracciones también son las frecuencias relativas de los diferentes valores de X en nuestro experimento. Entonces, realmente podemos calcular la media, o el promedio de un conjunto de datos, si conocemos los distintos valores que ocurren y sus frecuencias relativas sin tener conocimiento del número total de observaciones en el conjunto de datos. Por lo tanto, si 4/16 o 1/4 de los lanzamientos dan como resultado cero caras, 7/16 de los lanzamientos dan como resultado una cara y 5/16 dan como resultado dos caras, el número medio de caras por lanzamiento sería 1.06, sin importar si el número total de lanzamientos fue 16, 1000 o incluso 10,000.

Este método de frecuencias relativas se utiliza para calcular el número promedio de caras que esperaríamos obtener a largo plazo por el lanzamiento de dos monedas. A este valor promedio se le conoce como **media de la variable aleatoria** X o **media de la distribución de probabilidad de** X, y se le denota como μ_x o simplemente como μ cuando es evidente a qué variable aleatoria se está haciendo referencia. También es común entre los estadísticos referirse a esta media como la esperanza matemática o el valor esperado de la variable aleatoria X y denotarla como $E(X)$.

Suponiendo que una moneda legal se lanza dos veces, encontramos que el espacio muestral para el experimento es

$$S = \{HH, HT, TH, TT\}.$$

Como los 4 puntos muestrales son igualmente probables, se deduce que

$$P(X = 0) = P(TT) = \frac{1}{4}, \quad P(X = 1) = P(TH) + P(HT) = \frac{1}{2},$$

y

$$P(X = 2) = P(HH) = \frac{1}{4},$$

donde un elemento típico, digamos *TH*, indica que el primer lanzamiento dio como resultado una cruz seguida por una cara en el segundo lanzamiento. Así, estas probabilidades son precisamente las frecuencias relativas para los eventos dados a largo plazo. Por lo tanto,

$$\mu = E(X) = (0)\left(\frac{1}{4}\right) + (1)\left(\frac{1}{2}\right) + (2)\left(\frac{1}{4}\right) = 1.$$

Este resultado significa que una persona que lance 2 monedas una y otra vez obtendrá, en promedio, 1 cara por cada lanzamiento.

El método descrito antes para calcular el número esperado de caras cada vez que se lanzan 2 monedas sugiere que la media, o el valor esperado de cualquier variable aleatoria discreta, se puede obtener multiplicando cada uno de los valores $x_1, x_2, \ldots, x_n$ de la variable aleatoria X por su probabilidad correspondiente $f(x_1), f(x_2), \ldots, f(x_n)$ y sumando los productos. Esto es cierto, sin embargo, sólo si la variable aleatoria es discreta. En el caso de variables aleatorias continuas la definición de un valor esperado es esencialmente la misma, pero las sumatorias se reemplazan con integrales.

Definición 4.1: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La **media** o **valor esperado** de X es

$$\mu = E(X) = \sum_x xf(x)$$

si X es discreta, y

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

si X es continua.

El lector debe advertir que la forma para calcular el valor esperado, o media, que se muestra aquí es diferente del método para calcular la media muestral que se describió en el capítulo 1, donde la media muestral se obtuvo usando los datos. En la esperanza matemática el valor esperado se obtiene usando la distribución de probabilidad.

Sin embargo, la media suele considerarse un valor "central" de la distribución subyacente si se utiliza el valor esperado, como en la definición 4.1.

Ejemplo 4.1: Un inspector de calidad obtiene una muestra de un lote que contiene 7 componentes; el lote contiene 4 componentes buenos y 3 defectuosos. El inspector toma una muestra de 3 componentes. Calcule el valor esperado del número de componentes buenos en esta muestra.

***Solución*:** Sea X el número de componentes buenos en la muestra. La distribución de probabilidad de X es

$$f(x) = \frac{\binom{4}{x}\binom{3}{3-x}}{\binom{7}{3}}, \qquad x = 0, 1, 2, 3.$$

Unos cálculos sencillos dan $f(0) = 1/35$, $f(1) = 12/35$, $f(2) = 18/35$ y $f(3) = 4/35$. Por lo tanto,

$$\mu = E(X) = (0)\left(\frac{1}{35}\right) + (1)\left(\frac{12}{35}\right) + (2)\left(\frac{18}{35}\right) + (3)\left(\frac{4}{35}\right) = \frac{12}{7} = 1.7.$$

De esta manera, si de un lote de 4 componentes buenos y 3 defectuosos, se seleccionara al azar, una y otra vez, una muestra de tamaño 3, ésta contendría en promedio 1.7 componentes buenos. ∎

Ejemplo 4.2: Cierto día un vendedor de una empresa de aparatos médicos tiene dos citas. Considera que en la primera cita tiene 70 por ciento de probabilidades de cerrar una venta, por la cual podría obtener una comisión de $1000. Por otro lado, cree que en la segunda cita sólo tiene 40 por ciento de probabilidades de cerrar el trato, del cual obtendría $1500 de comisión. ¿Cuál es su comisión esperada con base en dichas probabilidades? Suponga que los resultados de las citas son independientes.

***Solución*:** En primer lugar sabemos que el vendedor, en las dos citas, puede obtener 4 comisiones totales: $0, $1000, $1500 y $2500. Necesitamos calcular sus probabilidades asociadas. Mediante la independencia obtenemos

$$f(\$0) = (1-0.7)(1-0.4) = 0.18, \quad f(\$2500) = (0.7)(0.4) = 0.28,$$
$$f(\$1000) = (0.7)(1-0.4) = 0.42, \text{ y } \quad f(\$1500) = (1-0.7)(0.4) = 0.12.$$

Por lo tanto, la comisión esperada para el vendedor es

$$\begin{aligned} E(X) =& (\$0)(0.18) + (\$1000)(0.42) + (\$1500)(0.12) + (\$2500)(0.28) \\ =& \$1300. \end{aligned}$$

∎

Los ejemplos 4.1 y 4.2 se diseñaron para que el lector comprenda mejor lo que queremos decir con la frase valor esperado de una variable aleatoria. En ambos casos las variables aleatorias son discretas. Seguimos con un ejemplo de variable aleatoria continua, donde un ingeniero se interesa en la *vida media* de cierto tipo de dispositivo electrónico. Ésta es una ilustración del problema *tiempo que transcurre antes de que se presente una falla* que se enfrenta a menudo en la práctica. El valor esperado de la vida del dispositivo es un parámetro importante para su evaluación.

Ejemplo 4.3: Sea X la variable aleatoria que denota la vida en horas de cierto dispositivo electrónico. La función de densidad de probabilidad es

$$f(x) = \begin{cases} \frac{20,000}{x^3}, & x > 100, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la vida esperada para esta clase de dispositivo.

***Solución*:** Si usamos la definición 4.1, tenemos

$$\mu = E(X) = \int_{100}^{\infty} x \frac{20,000}{x^3}\, dx = \int_{100}^{\infty} \frac{20,000}{x^2}\, dx = 200.$$

Por lo tanto, esperamos que este tipo de dispositivo dure *en promedio* 200 horas. ∎

Consideremos ahora una nueva variable aleatoria $g(X)$, la cual depende de X; es decir, cada valor de $g(X)$ es determinado por el valor de X. Por ejemplo, $g(X)$ podría ser X^2 o $3X - 1$, y siempre que X asuma el valor 2, $g(X)$ toma el valor $g(2)$. En particular, si X es una variable aleatoria discreta con distribución de probabilidad $f(x)$, para $x = -1, 0, 1, 2$ y $g(X) = X^2$, entonces,

$$\begin{aligned} P[g(X) = 0] &= P(X = 0) = f(0), \\ P[g(X) = 1] &= P(X = -1) + P(X = 1) = f(-1) + f(1), \\ P[g(X) = 4] &= P(X = 2) = f(2), \end{aligned}$$

así que la distribución de probabilidad de $g(X)$ se escribe como

$g(x)$	0	1	4
$P[g(X) = g(x)]$	$f(0)$	$f(-1) + f(1)$	$f(2)$

Por medio de la definición del valor esperado de una variable aleatoria obtenemos

$$\begin{aligned} \mu_{g(X)} &= E[g(x)] = 0f(0) + 1[f(-1) + f(1)] + 4f(2) \\ &= (-1)^2 f(-1) + (0)^2 f(0) + (1)^2 f(1) + (2)^2 f(2) = \sum_x g(x) f(x). \end{aligned}$$

Este resultado se generaliza en el teorema 4.1 para variables aleatorias discretas y continuas.

Teorema 4.1: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. El valor esperado de la variable aleatoria $g(X)$ es

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x) f(x)$$

si X es discreta, y

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

si X es continua.

Ejemplo 4.4: Suponga que el número de automóviles X que pasa por un local de lavado de autos entre las 4:00 P.M. y las 5:00 P.M. de cualquier viernes soleado tiene la siguiente distribución de probabilidad:

x	4	5	6	7	8	9
$P(X = x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

Sea $g(X) = 2X - 1$ la cantidad de dinero en dólares que el administrador paga al operador. Calcule las ganancias esperadas del operador en este periodo específico.

***Solución*:** Por el teorema 4.1, el operador puede esperar recibir

$$\begin{aligned} E[g(X)] = E(2X - 1) &= \sum_{x=4}^{9} (2x - 1) f(x) \\ &= (7)\left(\frac{1}{12}\right) + (9)\left(\frac{1}{12}\right) + (11)\left(\frac{1}{4}\right) + (13)\left(\frac{1}{4}\right) \\ &\quad + (15)\left(\frac{1}{6}\right) + (17)\left(\frac{1}{6}\right) = \$12.67. \end{aligned}$$

Ejemplo 4.5: Sea X una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado de $g(X) = 4X + 3$.

***Solución*:** Por el teorema 4.1 tenemos

$$E(4X + 3) = \int_{-1}^{2} \frac{(4x + 3)x^2}{3}\, dx = \frac{1}{3}\int_{-1}^{2} (4x^3 + 3x^2)\, dx = 8.$$

Debemos extender ahora nuestro concepto de esperanza matemática al caso de dos variables aleatorias X y Y con distribución de probabilidad conjunta $f(x, y)$.

Definición 4.2: Sean X y Y variables aleatorias con distribución de probabilidad conjunta $f(x, y)$. La media o valor esperado de la variable aleatoria $g(X, Y)$ es

$$\mu_{g(X,\,Y)} = E[g(X,\,Y)] = \sum_{x}\sum_{y} g(x, y) f(x, y)$$

si X y Y son discretas, y

$$\mu_{g(X,\,Y)} = E[g(X,\,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy$$

si X y Y son continuas.

Es evidente la generalización de la definición 4.2 para el cálculo de la esperanza matemática de funciones de varias variables aleatorias.

Ejemplo 4.6: Sean X y Y variables aleatorias con la distribución de probabilidad conjunta que se indica en la tabla 3.1 de la página 96. Calcule el valor esperado de $g(X, Y) = XY$. Por conveniencia se repite aquí la tabla.

	$f(x, y)$	x = 0	x = 1	x = 2	Totales por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

***Solución*:** Por la definición 4.2, escribimos

$$\begin{aligned} E(XY) &= \sum_{x=0}^{2}\sum_{y=0}^{2} xy f(x, y) \\ &= (0)(0) f(0, 0) + (0)(1) f(0, 1) \\ &\quad + (1)(0) f(1, 0) + (1)(1) f(1, 1) + (2)(0) f(2, 0) \\ &= f(1, 1) = \frac{3}{14}. \end{aligned}$$

Ejemplo 4.7: Calcule $E(Y/X)$ para la siguiente función de densidad

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

***Solución*:** Tenemos

$$E\left(\frac{Y}{X}\right) = \int_0^1 \int_0^2 \frac{y(1 + 3y^2)}{4}\, dxdy = \int_0^1 \frac{y + 3y^3}{2}\, dy = \frac{5}{8}.$$

Observe que si $g(X, Y) = X$ en la definición 4.2, tenemos

$$E(X) = \begin{cases} \sum_x \sum_y x f(x, y) = \sum_x x g(x) & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dy\, dx = \int_{-\infty}^{\infty} x g(x)\, dx & \text{(caso continuo)}, \end{cases}$$

donde $g(x)$ es la distribución marginal de X. Por lo tanto, para calcular $E(X)$ en un espacio bidimensional, se puede utilizar tanto la distribución de probabilidad conjunta de X y Y, como la distribución marginal de X. De manera similar, definimos

$$E(Y) = \begin{cases} \sum_y \sum_x y f(x, y) = \sum_y y h(y) & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y)\, dxdy = \int_{-\infty}^{\infty} y h(y)\, dy & \text{(caso continuo)}, \end{cases}$$

donde $h(y)$ es la distribución marginal de la variable aleatoria Y.

Ejercicios

4.1 En el ejercicio 3.13 de la página 92 se presenta la siguiente distribución de probabilidad de X, el número de imperfecciones que hay en cada 10 metros de una tela sintética, en rollos continuos de ancho uniforme

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Calcule el número promedio de imperfecciones que hay en cada 10 metros de esta tela.

4.2 La distribución de probabilidad de la variable aleatoria discreta X es

$$f(x) = \binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3.$$

Calcule la media de X.

4.3 Calcule la media de la variable aleatoria T que representa el total de las tres monedas del ejercicio 3.25 de la página 93.

4.4 Una moneda está cargada de manera que la probabilidad de ocurrencia de una cara es tres veces mayor que la de una cruz. Calcule el número esperado de cruces si esta moneda se lanza dos veces.

4.5 En un juego de azar a una mujer se le pagan \$3 si saca una jota o una reina, y \$5 si saca un rey o un as de una baraja ordinaria de 52 cartas. Si saca cualquier otra carta, pierde. ¿Cuánto debería pagar si el juego es justo?

4.6 A un operador de un local de lavado de autos se le paga de acuerdo con el número de automóviles que lava. Suponga que las probabilidades de que entre las 4:00 p.m. y las 5:00 p.m. de cualquier viernes soleado reciba \$7, \$9, \$11, \$13, \$15 o \$17 son: 1/12, 1/12, 1/4, 1/4, 1/6 y 1/6, respectivamente. Calcule las ganancias esperadas del operador para este periodo específico.

4.7 Si una persona invierte en unas acciones en particular, en un año tiene una probabilidad de 0.3 de obtener una ganancia de \$4000 o una probabilidad de 0.7 de tener una pérdida de \$1000. ¿Cuál es la ganancia esperada de esta persona?

4.8 Suponga que un distribuidor de joyería antigua está interesado en comprar un collar de oro para el que tiene 0.22 de probabilidades de venderlo con \$250 de utilidad; 0.36 de venderlo con \$150 de utilidad; 0.28 de venderlo al costo y 0.14 de venderlo con una pérdida de \$150. ¿Cuál es su utilidad esperada?

4.9 Un piloto privado desea asegurar su avión por \$200,000. La aseguradora estima que la probabilidad de pérdida total es de 0.002, que la probabilidad de una pérdida del 50% es de 0.01 y la probabilidad de una pérdida del 25% es de 0.1. Si se ignoran todas las demás pérdidas parciales, ¿qué prima debería cobrar cada año la aseguradora para tener una utilidad promedio de \$500?

4.10 Dos expertos en calidad de neumáticos examinan lotes de éstos y asignan a cada neumático puntuaciones de calidad en una escala de tres puntos. Sea X la puntuación dada por el experto A y Y la dada por el experto B. La siguiente tabla presenta la distribución conjunta para X y Y.

		y		
$f(x, y)$		1	2	3
x	1	0.10	0.05	0.02
	2	0.10	0.35	0.05
	3	0.03	0.10	0.20

Calcule μ_X y μ_Y.

4.11 La función de densidad de las mediciones codificadas del diámetro de paso de los hilos de un encaje es

$$f(x) = \begin{cases} \frac{4}{\pi(1+x^2)}, & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado de X.

4.12 Si la utilidad para un distribuidor de un automóvil nuevo, en unidades de \$5000, se puede ver como una variable aleatoria X que tiene la siguiente función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la utilidad promedio por automóvil.

4.13 La función de densidad de la variable aleatoria continua X, el número total de horas que una familia utiliza una aspiradora durante un año, en unidades de 100 horas, se da en el ejercicio 3.7 de la página 92 como

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2-x, & 1 \le x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el número promedio de horas por año que las familias utilizan sus aspiradoras.

4.14 Calcule la proporción X de personas que se podría esperar que respondieran a cierta encuesta que se envía por correo, si X tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{2(x+2)}{5}, & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

4.15 Suponga que dos variables aleatorias (X,Y) están distribuidas de manera uniforme en un círculo con radio a. Entonces, la función de densidad de probabilidad conjunta es

$$f(x, y) = \begin{cases} \frac{1}{\pi a^2}, & x^2 + y^2 \le a^2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule μ_X, el valor esperado de X.

4.16 Suponga que usted inspecciona un lote de 1000 bombillas de luz, entre las cuales hay 20 defectuosas, y elige al azar dos bombillas del lote sin reemplazo. Sean

$$X_1 = \begin{cases} 1, \text{ si la primera bombilla está defectuosa,} \\ 0, \text{ en otro caso.} \end{cases}$$

$$X_2 = \begin{cases} 1, \text{ si la segunda bombilla está defectuosa,} \\ 0, \text{ en otro caso.} \end{cases}$$

Calcule la probabilidad de que al menos una de las bombillas elegidas esté defectuosa. [*Sugerencia*: Calcule $P(X_1 + X_2 = 1)$.]

4.17 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	-3	6	9
$f(x)$	1/6	1/2	1/3

Calcule $\mu_{g(X)}$, donde $g(X) = (2X + 1)^2$.

4.18 Calcule el valor esperado de la variable aleatoria $g(X) = X^2$, donde X tiene la distribución de probabilidad del ejercicio 4.2.

4.19 Una empresa industrial grande compra varios procesadores de textos nuevos al final de cada año; el número exacto depende de la frecuencia de reparaciones del año anterior. Suponga que el número de procesadores de textos, X, que se compran cada año tiene la siguiente distribución de probabilidad:

x	0	1	2	3
$f(x)$	1/10	3/10	2/5	1/5

Si el costo del modelo deseado es de \$1200 por unidad y al final del año la empresa obtiene un descuento de $50X^2$ dólares, ¿cuánto espera gastar esta empresa en nuevos procesadores de textos durante este año?

4.20 Una variable aleatoria continua X tiene la siguiente función de densidad

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado de $g(X) = e^{2X/3}$.

4.21 ¿Cuál es la utilidad promedio por automóvil que obtiene un distribuidor, si la utilidad en cada uno está dada por $g(X) = X^2$, donde X es una variable aleatoria que tiene la función de densidad del ejercicio 4.12?

4.22 El periodo de hospitalización, en días, para pacientes que siguen el tratamiento para cierto tipo de trastorno renal es una variable aleatoria $Y = X + 4$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el número promedio de días que una persona permanece hospitalizada con el fin de seguir el tratamiento para dicha enfermedad.

4.23 Suponga que X y Y tienen la siguiente función de probabilidad conjunta:

$f(x, y)$		x = 2	x = 4
y	1	0.10	0.15
	3	0.20	0.30
	5	0.10	0.15

a) Calcule el valor esperado de $g(X, Y) = XY^2$.
b) Calcule μ_X y μ_Y.

4.24 Remítase a las variables aleatorias cuya distribución de probabilidad conjunta se da en el ejercicio 3.39 de la página 105 y
a) calcule $E(X^2Y - 2XY)$;
b) calcule $\mu_X - \mu_Y$.

4.25 Remítase a las variables aleatorias cuya distribución de probabilidad conjunta se da en el ejercicio 3.51 de la página 106 y calcule la media para el número total de jotas y reyes cuando se sacan 3 cartas, sin reemplazo, de las 12 cartas mayores de una baraja ordinaria de 52 cartas.

4.26 Sean X y Y las siguientes variables aleatorias con función de densidad conjunta

$$f(x, y) = \begin{cases} 4xy, & 0 < x,\ y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado de $Z = \sqrt{X^2 + Y^2}$.

4.27 En el ejercicio 3.27 de la página 93 una función de densidad está dada por el tiempo que tarda en fallar un componente importante de un reproductor de DVD. Calcule el número medio de horas antes de que empiece a fallar el componente y, por lo tanto, el reproductor de DVD.

4.28 Considere la información del ejercicio 3.28 de la página 93. El problema tiene que ver con el peso, en onzas, del producto que contiene una caja de cereal con

$$f(x) = \begin{cases} \frac{2}{5}, & 23.75 \le x \le 26.25, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Grafique la función de densidad.
b) Calcule el valor esperado o peso medio en onzas.
c) ¿Se sorprende de su respuesta en *b*)? Explique lo que responda.

4.29 El ejercicio 3.29 de la página 93 se refiere a una importante distribución del tamaño de las partículas caracterizada por

$$f(x) = \begin{cases} 3x^{-4}, & x > 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Grafique la función de densidad.
b) Determine el tamaño medio de la partícula.

4.30 En el ejercicio 3.31 de la página 94 la distribución del tiempo que transcurre antes de que una lavadora requiera una reparación mayor fue dada como

$$f(y) = \begin{cases} \frac{1}{4}e^{-y/4}, & y \geq 0, \\ 0, & \text{en otro caso.} \end{cases}$$

¿Cuál es la media de población del tiempo que transcurre antes de requerir la reparación?

4.31 Considere el ejercicio 3.32 de la página 94.
a) ¿Cuál es la proporción media del presupuesto asignado para el control ambiental y de la contaminación?
b) ¿Cuál es la probabilidad de que una empresa elegida al azar tenga una proporción asignada para el control ambiental y de la contaminación que exceda la media de la población dada en *a*)?

4.32 En el ejercicio 3.13 de la página 92 la distribución del número de imperfecciones en cada 10 metros de tela sintética fue dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

a) Grafique la función de probabilidad.
b) Calcule el número de imperfecciones esperado $E(X) = \mu$.
c) Calcule $E(X^2)$.

4.2 Varianza y covarianza de variables aleatorias

La media o valor esperado de una variable aleatoria X es de especial importancia en estadística porque describe en dónde se centra la distribución de probabilidad. Sin embargo, la media por sí misma no ofrece una descripción adecuada de la forma de la distribución. También se necesita clasificar la variabilidad en la distribución. En la figura 4.1 tenemos los histogramas de dos distribuciones de probabilidad discretas con la misma media $\mu = 2$, pero que difieren de manera considerable en la variabilidad o dispersión de sus observaciones sobre la media.

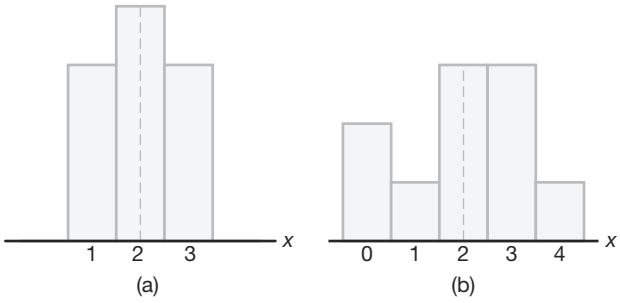


Figura 4.1: Distribuciones con medias iguales y dispersiones diferentes.

La medida de variabilidad más importante de una variable aleatoria X se obtiene aplicando el teorema 4.1 con $g(X) = (X - \mu)^2$. A esta cantidad se le denomina **varianza de la variable aleatoria** X o **varianza de la distribución de probabilidad de** X y se

denota como Var(X), o con el símbolo σ_X^2, o simplemente como σ^2 cuando es evidente a qué variable aleatoria se está haciendo referencia.

Definición 4.3: Sea X una variable aleatoria con distribución de probabilidad $f(x)$ y media μ. La varianza de X es

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x), \quad \text{si } X \text{ es discreta, y}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \quad \text{si } X \text{ es continua.}$$

La raíz cuadrada positiva de la varianza, σ, se llama **desviación estándar** de X.

La cantidad $x - \mu$ en la definición 4.3 se llama **desviación de una observación** respecto a su media. Como estas desviaciones se elevan al cuadrado y después se promedian, σ^2 será mucho menor para un conjunto de valores x que estén cercanos a μ, que para un conjunto de valores que varíe de forma considerable de μ.

Ejemplo 4.8: Suponga que la variable aleatoria X representa el número de automóviles que se utilizan con propósitos de negocios oficiales en un día de trabajo dado. La distribución de probabilidad para la empresa A [figura 4.1(a)] es

x	1	2	3
$f(x)$	0.3	0.4	0.3

y para la empresa B [figura 4.1(b)] es

x	0	1	2	3	4
$f(x)$	0.2	0.1	0.3	0.3	0.1

Demuestre que la varianza de la distribución de probabilidad para la empresa B es mayor que la de la empresa A.

***Solución*:** Para la empresa A encontramos que

$$\mu_A = E(X) = (1)(0.3) + (2)(0.4) + (3)(0.3) = 2.0,$$

y entonces

$$\sigma_A^2 = \sum_{x=1}^{3} (x-2)^2 = (1-2)^2(0.3) + (2-2)^2(0.4) + (3-2)^2(0.3) = 0.6.$$

Para la empresa B tenemos

$$\mu_B = E(X) = (0)(0.2) + (1)(0.1) + (2)(0.3) + (3)(0.3) + (4)(0.1) = 2.0,$$

y entonces

$$\begin{aligned}\sigma_B^2 &= \sum_{x=0}^{4} (x-2)^2 f(x) \\ &= (0-2)^2(0.2) + (1-2)^2(0.1) + (2-2)^2(0.3) \\ &\quad + (3-2)^2(0.3) + (4-2)^2(0.1) = 1.6.\end{aligned}$$

Es evidente que la varianza del número de automóviles que se utilizan con propósitos de negocios oficiales es mayor para la empresa *B* que para la empresa *A*.

Una fórmula alternativa que se prefiere para calcular σ^2, que a menudo simplifica los cálculos, se establece en el siguiente teorema.

Teorema 4.2: La varianza de una variable aleatoria *X* es

$$\sigma^2 = E(X^2) - \mu^2.$$

***Prueba*:** Para el caso discreto escribimos

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) \\ &= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x).\end{aligned}$$

Como $\mu = \sum_x x f(x)$ por definición, y $\sum_x f(x) = 1$ para cualquier distribución de probabilidad discreta, se deduce que

$$\sigma^2 = \sum x^2 f(x) - \mu^2 = E(X^2) - \mu^2.$$

Para el caso continuo la demostración es la misma paso a paso, reemplazando las sumatorias por integrales.

Ejemplo 4.9: Suponga que la variable aleatoria *X* representa el número de partes defectuosas de una máquina cuando de una línea de producción se obtiene una muestra de tres partes y se somete a prueba. La siguiente es la distribución de probabilidad de *X*.

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

Utilice el teorema 4.2 y calcule σ^2.

***Solución*:** Primero calculamos

$$\mu = (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) = 0.61.$$

Luego,

$$E(X^2) = (0)(0.51) + (1)(0.38) + (4)(0.10) + (9)(0.01) = 0.87.$$

Por lo tanto,

$$\sigma^2 = 0.87 - (0.61)^2 = 0.4979.$$

Ejemplo 4.10: La demanda semanal de una bebida para una cadena local de tiendas de abarrotes, en miles de litros, es una variable aleatoria continua *X* que tiene la siguiente densidad de probabilidad

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la media y la varianza de *X*.

***Solución*:** Al calcular $E(X)$ y $E(X^2)$ tenemos

$$\mu = E(X) = 2\int_1^2 x(x-1)\,dx = \frac{5}{3}$$

y

$$E(X^2) = 2\int_1^2 x^2(x-1)\,dx = \frac{17}{6}.$$

Por lo tanto,

$$\sigma^2 = \frac{17}{6} - \left(\frac{5}{3}\right)^2 = \frac{1}{18}.$$

Hasta el momento la varianza o la desviación estándar sólo tiene significado cuando comparamos dos o más distribuciones que tienen las mismas unidades de medida. Por lo tanto, podemos comparar las varianzas de las distribuciones de contenido, medido en litros, de botellas de jugo de naranja de dos empresas, y el valor más grande indicaría la empresa cuyo producto es más variable o menos uniforme. No tendría caso comparar la varianza de una distribución de estaturas con la varianza de una distribución de calificaciones de aptitud. En la sección 4.4 mostramos cómo se utiliza la desviación estándar para describir una sola distribución de observaciones.

Extenderemos ahora nuestro concepto de varianza de una variable aleatoria X para incluir también variables aleatorias relacionadas con X. Para la variable aleatoria $g(X)$ la varianza se denotará por $\sigma^2_{g(X)}$ y se calculará empleando el siguiente teorema.

Teorema 4.3: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La varianza de la variable aleatoria $g(X)$ es

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \sum_x [g(x) - \mu_{g(X)}]^2 f(x)$$

si X es discreta, y

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 f(x)\,dx$$

si X es continua.

***Prueba*:** Como $g(X)$ es en sí misma una variable aleatoria con media $\mu_{g(X)}$, como se define en el teorema 4.1, de la definición 4.3 se deduce que

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]\}.$$

Ahora bien, la demostración se completa aplicando nuevamente el teorema 4.1 a la variable aleatoria $[g(X) - \mu_{g(X)}]^2$.

Ejemplo 4.11: Calcule la varianza de $g(X) = 2X + 3$, donde X es una variable aleatoria con la siguiente distribución de probabilidad

x	0	1	2	3
$f(x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

***Solución*:** Primero se calcula la media de la variable aleatoria $2X + 3$. De acuerdo con el teorema 4.1,

$$\mu_{2X+3} = E(2X + 3) = \sum_{x=0}^{3} (2x + 3) f(x) = 6.$$

Ahora, usando el teorema 4.3, tenemos

$$\sigma^2_{2X+3} = E\{[(2X + 3) - \mu_{2x+3}]^2\} = E[(2X + 3 - 6)^2]$$
$$= E(4X^2 - 12X + 9) = \sum_{x=0}^{3} (4x^2 - 12x + 9) f(x) = 4.$$

Ejemplo 4.12: Sea X una variable aleatoria que tiene la función de densidad dada en el ejemplo 4.5 de la página 115. Calcule la varianza de la variable aleatoria $g(X) = 4X + 3$.

***Solución*:** En el ejemplo 4.5 encontramos que $\mu_{4X+3} = 8$. Ahora bien, usando el teorema 4.3,

$$\sigma^2_{4X+3} = E\{[(4X + 3) - 8]^2\} = E[(4X - 5)^2]$$
$$= \int_{-1}^{2} (4x - 5)^2 \frac{x^2}{3}\, dx = \frac{1}{3}\int_{-1}^{2} (16x^4 - 40x^3 + 25x^2)\, dx = \frac{51}{5}.$$

Si $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$, donde $\mu_X = E(X)$ y $\mu_Y = E(Y)$, la definición 4.2 da un valor esperado denominado **covarianza** de X y Y, que se denota por σ_{XY} o Cov(X, Y).

Definición 4.4: Sean X y Y variables aleatorias con distribución de probabilidad conjunta $f(x, y)$. La covarianza de X y Y es

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_y) f(x, y)$$

si X y Y son discretas, y

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_y) f(x, y)\, dx\, dy$$

si X y Y son continuas.

La covarianza entre dos variables aleatorias es una medida de la naturaleza de la asociación entre ambas. Si valores grandes de X a menudo dan como resultado valores grandes de Y, o valores pequeños de X, dan como resultado valores pequeños de Y, $X - \mu_X$ positiva con frecuencia dará como resultado $Y - \mu_Y$ positiva, y $X - \mu_X$ negativa a menudo dará como resultado $Y - \mu_Y$ negativa. Por consiguiente, el producto $(X - \mu_X)(Y - \mu_Y)$ tenderá a ser positivo. Por otro lado, si con frecuencia valores grandes de X dan como resultado valores pequeños de Y, entonces el producto $(X - \mu_X)(Y - \mu_Y)$ tenderá a ser negativo. El *signo* de la covarianza indica si la relación entre dos variables aleatorias dependientes es positiva o negativa. Cuando X y Y son estadísticamente independientes, se puede demostrar que la covarianza es cero (véase el corolario 4.5). Lo opuesto, sin embargo, por lo general no es cierto. Dos variables pueden tener covarianza cero y aun así no ser estadísticamente independientes. Observe que la covarianza sólo describe la relación *lineal* entre dos variables aleatorias. Por consiguiente, si una covarianza entre X y Y es cero, X y Y podrían tener una relación no lineal, lo cual significa que no necesariamente son independientes.

La fórmula alternativa que se prefiere para σ_{XY} se establece en el teorema 4.4.

Teorema 4.4: La covarianza de dos variables aleatorias X y Y, con medias μ_X y μ_Y, respectivamente, está dada por

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y .$$

***Prueba*:** Para el caso discreto escribimos

$$\begin{aligned}\sigma_{XY} &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ &= \sum_x \sum_y xy\, f(x, y) - \mu_X \sum_x \sum_y y\, f(x, y) \\ &\quad - \mu_Y \sum_x \sum_y x f(x, y) + \mu_X \mu_Y \sum_x \sum_y f(x, y).\end{aligned}$$

Dado que

$$\mu_X = \sum_x x f(x, y), \quad \mu_Y = \sum_y y f(x, y), \text{ y } \sum_x \sum_y f(x, y) = 1$$

para cualquier distribución discreta conjunta se deduce que

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y.$$

Para el caso continuo la demostración es idéntica, pero las sumatorias se reemplazan por integrales. ∎

Ejemplo 4.13: En el ejemplo 3.14 de la página 95 se describe una situación acerca del número de repuestos azules X y el número de repuestos rojos Y. Cuando de cierta caja se seleccionan dos repuestos para bolígrafo al azar y la distribución de probabilidad conjunta es la siguiente,

		x			
	$f(x, y)$	0	1	2	$h(y)$
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
	$g(x)$	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Calcule la covarianza de X y Y.

***Solución*:** Del ejemplo 4.6 vemos que $E(XY) = 3/14$. Ahora bien,

$$\mu_X = \sum_{x=0}^{2} x g(x) = (0)\left(\frac{5}{14}\right) + (1)\left(\frac{15}{28}\right) + (2)\left(\frac{3}{28}\right) = \frac{3}{4},$$

y

$$\mu_Y = \sum_{y=0}^{2} y h(y) = (0)\left(\frac{15}{28}\right) + (1)\left(\frac{3}{7}\right) + (2)\left(\frac{1}{28}\right) = \frac{1}{2}.$$

Por lo tanto,

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{3}{14} - \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = -\frac{9}{56}.$$

Ejemplo 4.14: La fracción X de corredores y la fracción Y de corredoras que compiten en carreras de maratón se describen mediante la función de densidad conjunta

$$f(x, y) = \begin{cases} 8xy, & 0 \le y \le x \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la covarianza de X y Y.

***Solución*:** Primero calculamos las funciones de densidad marginal. Éstas son

$$g(x) = \begin{cases} 4x^3, & 0 \le x \le 1, \\ 0, & \text{en otro caso,} \end{cases}$$

y

$$h(y) = \begin{cases} 4y(1-y^2), & 0 \le y \le 1, \\ 0, & \text{en otro caso,} \end{cases}$$

A partir de las funciones de densidad marginal dadas, calculamos

$$\mu_X = E(X) = \int_0^1 4x^4 \, dx = \frac{4}{5} \quad \text{y} \quad \mu_Y = \int_0^1 4y^2(1-y^2) \, dy = \frac{8}{15}.$$

De las funciones de densidad conjunta dadas arriba, tenemos

$$E(XY) = \int_0^1 \int_y^1 8x^2y^2 \, dx \, dy = \frac{4}{9}.$$

Entonces,

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}.$$

Aunque la covarianza entre dos variables aleatorias brinda información respecto de la naturaleza de la relación, la magnitud de σ_{XY} *no indica nada respecto a la fuerza de la relación,* ya que σ_{XY} depende de la escala. Su magnitud dependerá de las unidades que se utilicen para medir X y Y. Hay una versión de la covarianza sin escala que se denomina **coeficiente de correlación** y se utiliza ampliamente en estadística.

Definición 4.5: Sean X y Y variables aleatorias con covarianza σ_{XY} y desviaciones estándar σ_X y σ_Y, respectivamente. El coeficiente de correlación de X y Y es

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Debería quedar claro para el lector que ρ_{XY} no tiene las unidades de X y Y. El coeficiente de correlación satisface la desigualdad $-1 \le \rho_{XY} \le 1$. Toma un valor de cero cuando $\sigma_{XY} = 0$. Donde hay una dependencia lineal exacta, digamos $Y \equiv a + bX$, $\rho_{XY} = 1$ si

$b > 0$ y $\rho_{XY} = -1$ si $b < 0$. (Véase el ejercicio 4.48). En el capítulo 12, donde examinaremos la regresión lineal, analizamos más a fondo el coeficiente de correlación.

Ejemplo 4.15: Calcule el coeficiente de correlación entre X y Y en el ejemplo 4.13.
***Solución*:** Dado que

$$E(X^2) = (0^2)\left(\frac{5}{14}\right) + (1^2)\left(\frac{15}{28}\right) + (2^2)\left(\frac{3}{28}\right) = \frac{27}{28}$$

y

$$E(Y^2) = (0^2)\left(\frac{15}{28}\right) + (1^2)\left(\frac{3}{7}\right) + (2^2)\left(\frac{1}{28}\right) = \frac{4}{7},$$

obtenemos

$$\sigma_X^2 = \frac{27}{28} - \left(\frac{3}{4}\right)^2 = \frac{45}{112} \quad \text{y} \quad \sigma_Y^2 = \frac{4}{7} - \left(\frac{1}{2}\right)^2 = \frac{9}{28}.$$

Por lo tanto, el coeficiente de correlación entre X y Y es

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y} = \frac{-9/56}{\sqrt{(45/112)(9/28)}} = -\frac{1}{\sqrt{5}}.$$

Ejemplo 4.16: Calcule el coeficiente de correlación entre X y Y en el ejemplo 4.14.
***Solución*:** Dado que

$$E(X^2) = \int_0^1 4x^5\,dx = \frac{2}{3} \quad \text{y} \quad E(Y^2) = \int_0^1 4y^3(1-y^2)\,dy = 1 - \frac{2}{3} = \frac{1}{3},$$

concluimos que

$$\sigma_X^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75} \quad \text{y} \quad \sigma_Y^2 = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = \frac{11}{225}.$$

Por lo tanto,

$$\rho_{XY} = \frac{4/225}{\sqrt{(2/75)(11/225)}} = \frac{4}{\sqrt{66}}.$$

Observe que, aunque la covarianza en el ejemplo 4.15 tiene mayor magnitud (sin importar el signo) que la del ejemplo 4.16, la relación entre las magnitudes de los coeficientes de correlación en estos dos ejemplos es exactamente la inversa. Esto es evidencia de que no debemos basarnos en la magnitud de la covarianza para determinar la fuerza de la relación.

Ejercicios

4.33 Use la definición 4.3 de la página 120 para encontrar la varianza de la variable aleatoria X del ejercicio 4.7 de la página 117.

4.34 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	-2	3	5
$f(x)$	0.3	0.2	0.5

Calcule la desviación estándar de X.

4.35 La variable aleatoria X, que representa el número de errores por 100 líneas de código de programación, tiene la siguiente distribución de probabilidad:

x	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	0.3	0.04

Utilice el teorema 4.2 de la página 121 para calcular la varianza de X.

4.36 Suponga que las probabilidades de que 0, 1, 2 o 3 fallas de energía eléctrica afecten cierta subdivisión en cualquier año dado son 0.4, 0.3, 0.2 y 0.1, respectivamente. Calcule la media y la varianza de la variable aleatoria X que representa el número de fallas de energía que afectan esta subdivisión.

4.37 La utilidad que obtiene un distribuidor, en unidades de $5000, al vender un automóvil nuevo es una variable aleatoria X que tiene la función de densidad que se presenta en el ejercicio 4.12 de la página 117. Calcule la varianza de X.

4.38 La proporción de personas que responden cierta encuesta que se manda por correo es una variable aleatoria X, la cual tiene la función de densidad del ejercicio 4.14 de la página 117. Calcule la varianza de X.

4.39 El número total de horas que una familia utiliza una aspiradora en un año, en unidades de 100 horas, es una variable aleatoria X que tiene la función de densidad dada en el ejercicio 4.13 de la página 117. Calcule la varianza de X.

4.40 Remítase al ejercicio 4.14 de la página 117 y calcule $\sigma^2_{g(X)}$ para la función $g(X) = 3X^2 + 4$.

4.41 Calcule la desviación estándar de la variable aleatoria $g(X) = (2X + 1)^2$ del ejercicio 4.17 en la página 118.

4.42 Utilice los resultados del ejercicio 4.21 de la página 118 y calcule la varianza de $g(X) = X^2$, donde X es una variable aleatoria que tiene la función de densidad del ejercicio 4.12 de la página 117.

4.43 El tiempo que transcurre, en minutos, para que un avión obtenga vía libre para despegar en cierto aeropuerto es una variable aleatoria $Y = 3X - 2$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4}, & x > 0 \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la media y la varianza de la variable aleatoria Y.

4.44 Calcule la covarianza de las variables aleatorias X y Y del ejercicio 3.39 de la página 105.

4.45 Calcule la covarianza de las variables aleatorias X y Y del ejercicio 3.49 de la página 106.

4.46 Calcule la covarianza de las variables aleatorias X y Y del ejercicio 3.44 de la página 105.

4.47 Calcule la covarianza de las variables aleatorias X y Y cuya función de densidad conjunta está dada en el ejercicio 3.40 de la página 105.

4.48 Dada una variable aleatoria X, con desviación estándar σ_X y una variable aleatoria $Y = a + bX$, demuestre que si $b < 0$, el coeficiente de correlación $\rho_{XY} = -1$, y si $b > 0$, $\rho_{XY} = 1$.

4.49 Considere la situación del ejercicio 4.32 de la página 119. La distribución del número de imperfecciones por cada 10 metros de tela sintética está dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Calcule la varianza y la desviación estándar del número de imperfecciones.

4.50 En una tarea de laboratorio, si el equipo está funcionando, la función de densidad del resultado observado X es

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la varianza y la desviación estándar de X.

4.51 Determine el coeficiente de correlación entre X y Y para las variables aleatorias X y Y del ejercicio 3.39 de la página 105.

4.52 Las variables aleatorias X y Y tienen la siguiente distribución conjunta

$$f(x, y) = \begin{cases} 2, & 0 < x \le y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Determine el coeficiente de correlación entre X y Y.

4.3 Medias y varianzas de combinaciones lineales de variables aleatorias

Ahora estudiaremos algunas propiedades útiles que simplificarán los cálculos de las medias y las varianzas de variables aleatorias que aparecen en los siguientes capítulos. Estas propiedades nos permitirán ocuparnos de las esperanzas matemáticas en términos de otros parámetros que ya conocemos o que ya calculamos con facilidad. Todos los resultados que presentamos aquí son válidos para variables aleatorias continuas y discretas. Las demostraciones se dan sólo para el caso continuo. Comenzamos con un teorema y dos corolarios que deberían ser, de forma intuitiva, razonables para el lector.

Teorema 4.5: Si a y b son constantes, entonces,

$$E(aX + b) = aE(X) + b.$$

***Prueba*:** Por la definición de valor esperado,

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f(x)\, dx = a \int_{-\infty}^{\infty} x f(x)\, dx + b \int_{-\infty}^{\infty} f(x)\, dx.$$

La primera integral de la derecha es $E(X)$ y la segunda integral es igual a 1. Por lo tanto,

$$E(aX + b) = aE(X) + b.$$

Corolario 4.1: Al establecer que $a = 0$ vemos que $E(b) = b$.

Corolario 4.2: Al establecer que $b = 0$ vemos que $E(aX) = aE(X)$.

Ejemplo 4.17: Aplique el teorema 4.5 a la variable aleatoria discreta $f(X) = 2X - 1$ para resolver de nuevo el ejemplo 4.4 de la página 115.

***Solución*:** De acuerdo con el teorema 4.5, escribimos

$$E(2X - 1) = 2E(X) - 1.$$

Ahora,

$$\mu = E(X) = \sum_{x=4}^{9} x f(x)$$

$$= (4)\left(\frac{1}{12}\right) + (5)\left(\frac{1}{12}\right) + (6)\left(\frac{1}{4}\right) + (7)\left(\frac{1}{4}\right) + (8)\left(\frac{1}{6}\right) + (9)\left(\frac{1}{6}\right) = \frac{41}{6}.$$

Por lo tanto,

$$\mu_{2X-1} = (2)\left(\frac{41}{6}\right) - 1 = \$12.67,$$

como antes.

Ejemplo 4.18: Para resolver de nuevo el ejemplo 4.5 de la página 115 aplique el teorema 4.5 a la variable aleatoria continua $g(X) = 4X + 3$.

***Solución*:** En el ejemplo 4.5 utilizamos el teorema 4.5 para escribir

$$E(4X + 3) = 4E(X) + 3.$$

Ahora,

$$E(X) = \int_{-1}^{2} x\left(\frac{x^2}{3}\right) dx = \int_{-1}^{2} \frac{x^3}{3}\, dx = \frac{5}{4}.$$

Por lo tanto,

$$E(4X + 3) = (4)\left(\frac{5}{4}\right) + 3 = 8,$$

como antes.

Teorema 4.6: El valor esperado de la suma o diferencia de dos o más funciones de una variable aleatoria X es la suma o diferencia de los valores esperados de las funciones. Es decir,

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

***Prueba*:** Por definición,

$$\begin{aligned} E[g(X) \pm h(X)] &= \int_{-\infty}^{\infty} [g(x) \pm h(x)] f(x)\, dx \\ &= \int_{-\infty}^{\infty} g(x) f(x)\, dx \pm \int_{-\infty}^{\infty} h(x) f(x)\, dx \\ &= E[g(X)] \pm E[h(X)]. \end{aligned}$$

Ejemplo 4.19: Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

Calcule el valor esperado de $Y = (X - 1)^2$.

***Solución*:** Si aplicamos el teorema 4.6 a la función $Y = (X - 1)^2$, podemos escribir

$$E[(X - 1)^2] = E(X^2 - 2X + 1) = E(X^2) - 2E(X) + E(1).$$

A partir del corolario 4.1, $E(1) = 1$, y por cálculo directo

$$E(X) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (2)(0) + (3)\left(\frac{1}{6}\right) = 1 \text{ y}$$

$$E(X^2) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (4)(0) + (9)\left(\frac{1}{6}\right) = 2.$$

En consecuencia,

$$E[(X - 1)^2] = 2 - (2)(1) + 1 = 1.$$

Ejemplo 4.20: La demanda semanal de cierta bebida en una cadena de tiendas de abarrotes, en miles de litros, es una variable aleatoria continua $g(X) = X^2 + X - 2$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado para la demanda semanal de la bebida.

Solución: Por medio del teorema 4.6, escribimos

$$E(X^2 + X - 2) = E(X^2) + E(X) - E(2).$$

A partir del corolario 4.1, $E(2) = 2$, y por integración directa,

$$E(X) = \int_1^2 2x(x-1)\,dx = \frac{5}{3} \quad \text{y} \quad E(X^2) = \int_1^2 2x^2(x-1)\,dx = \frac{17}{6}.$$

Entonces,

$$E(X^2 + X - 2) = \frac{17}{6} + \frac{5}{3} - 2 = \frac{5}{2},$$

así que la demanda semanal promedio de la bebida en esta cadena de tiendas de abarrotes es de 2500 litros. ∎

Suponga que tenemos dos variables aleatorias X y Y con distribución de probabilidad conjunta $f(x, y)$. Dos propiedades adicionales que serán muy útiles en los capítulos siguientes incluyen los valores esperados de la suma, la diferencia y el producto de estas dos variables aleatorias. Sin embargo, comenzaremos por demostrar un teorema sobre el valor esperado de la suma o diferencia de funciones de las variables dadas. Por supuesto, tan sólo se trata de una extensión del teorema 4.6.

Teorema 4.7: El valor esperado de la suma o diferencia de dos o más funciones de las variables aleatorias X y Y es la suma o diferencia de los valores esperados de las funciones. Es decir,

$$E[g(X, Y) \pm h(X, Y)] = E[g(X, Y)] \pm E[h(X, Y)].$$

***Prueba*:** Por la definición 4.2,

$$\begin{aligned} E[g(X, Y) \pm h(X, Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [g(x, y) \pm h(x, y)] f(x, y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\,dx\,dy \pm \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y)\,dx\,dy \\ &= E[g(X, Y)] \pm E[h(X, Y)]. \end{aligned}$$

∎

Corolario 4.3: Si establecemos que $g(X, Y) = g(X)$ y $h(X, Y) = h(Y)$, vemos que

$$E[g(X) \pm h(Y)] = E[g(X)] \pm E[h(Y)].$$

Corolario 4.4: Si establecemos que $g(X, Y) = X$ y $h(X, Y) = Y$, vemos que

$$E[X \pm Y] = E[X] \pm E[Y].$$

Si X representa la producción diaria de algún artículo de la máquina A y Y la producción diaria del mismo artículo de la máquina B, entonces $X + Y$ representa la cantidad total de artículos que ambas máquinas producen diariamente. El corolario 4.4 establece que la producción diaria promedio para ambas máquinas es igual a la suma de la producción diaria promedio de cada máquina.

Teorema 4.8: Sean X y Y dos variables aleatorias independientes. Entonces,

$$E(XY) = E(X)E(Y).$$

***Prueba*:** Por la definición 4.2,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y)\, dx\, dy.$$

Como X y Y son independientes, podemos escribir

$$f(x, y) = g(x)h(y),$$

donde $g(x)$ y $h(y)$ son las distribuciones marginales de X y Y, respectivamente. En consecuencia,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, g(x)h(y)\, dx\, dy = \int_{-\infty}^{\infty} xg(x)\, dx \int_{-\infty}^{\infty} yh(y)\, dy \\ &= E(X)\, E(Y). \end{aligned}$$

Para variables discretas el teorema 4.8 se ilustra mediante un experimento en el que se lanzan un dado verde y uno rojo. La variable aleatoria X representa el resultado de lanzar el dado verde y la variable aleatoria Y el resultado de lanzar el dado rojo. Entonces XY representa el producto de los números que resultan de lanzar el par de dados. A la larga el promedio de los productos de los números es igual al producto del número promedio que resulta de lanzar el dado verde y el número promedio que resulta de lanzar el dado rojo.

Corolario 4.5: Sean X y Y dos variables aleatorias independientes. Entonces, $\sigma_{XY} = 0$.

***Prueba*:** La demostración se puede realizar utilizando los teoremas 4.4 y 4.8.

Ejemplo 4.21: Se sabe que la proporción de galio y arseniuro no afecta el funcionamiento de las obleas de arseniuro de galio que son los principales componentes de los circuitos integrados. Denotemos con X la proporción de galio a arseniuro y con Y el porcentaje de obleas funcionales producidas durante una hora. X y Y son variables aleatorias independientes con la siguiente función de densidad conjunta

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Demuestre que $E(XY) = E(X)E(Y)$, como sugiere el teorema 4.8.

***Solución*:** Por definición,

$$E(XY) = \int_0^1 \int_0^2 \frac{x^2 y(1+3y^2)}{4}\, dxdy = \frac{5}{6},\ E(X) = \frac{4}{3},\ \text{y} \quad E(Y) = \frac{5}{8}.$$

Por lo tanto,

$$E(X)E(Y) = \left(\frac{4}{3}\right)\left(\frac{5}{8}\right) = \frac{5}{6} = E(XY).$$

Concluimos esta sección con la demostración de un teorema y la presentación de varios corolarios que son útiles para calcular varianzas o desviaciones estándar.

Teorema 4.9: Si X y Y son variables aleatorias con distribución de probabilidad conjunta $f(x, y)$, y a, b y c son constantes, entonces

$$\sigma^2_{aX+bY+c} \quad a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}.$$

***Prueba*:** Por definición, $\sigma^2_{aX+bY+c} = E\{[(aX + bY + c) - \mu_{aX+bY+c}]^2\}$. Entonces,

$$\mu_{aX+bY+c} = E(aX + bY + c) = aE(X) + bE(Y) + c = a\mu_X + b\mu_Y + c,$$

si utilizamos el corolario 4.4 y después el corolario 4.2. Por lo tanto,

$$\begin{aligned}\sigma^2_{aX+bY+c} &= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\ &= a^2E[(X - \mu_X)^2] + b^2E[(Y - \mu_Y)^2] + 2abE[(X - \mu_X)(Y - \mu_Y)] \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}.\end{aligned}$$

Si utilizamos el teorema 4.9, tenemos los siguientes corolarios.

Corolario 4.6: Si se establece que $b = 0$, vemos que

$$\sigma^2_{aX+c} = a^2\sigma_X^2 = a^2\sigma^2.$$

Corolario 4.7: Si se establece que $a = 1$ y $b = 0$, vemos que

$$\sigma^2_{X+c} = \sigma_X^2 = \sigma^2.$$

Corolario 4.8: Si se establece que $b = 0$ y $c = 0$, vemos que

$$\sigma^2_{aX} = a^2\sigma_X^2 = a^2\sigma^2.$$

Los corolarios 4.6 y 4.7 establecen que la varianza no cambia si se suma o se resta una constante a una variable aleatoria. La suma o resta de una constante simplemente corre los valores de X a la derecha o a la izquierda, pero no cambia su variabilidad. Sin embargo, si una variable aleatoria se multiplica por una constante o se divide entre ésta, entonces los corolarios 4.6 y 4.8 establecen que la varianza se multiplica por el cuadrado de la constante o se divide entre éste.

Corolario 4.9: Si X y Y son variables aleatorias independientes, entonces

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

El resultado que se establece en el corolario 4.9 se obtiene a partir del teorema 4.9 y recurriendo al corolario 4.5.

Corolario 4.10: Si X y Y son variables aleatorias independientes, entonces,

$$\sigma^2_{aX-bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

El corolario 4.10 se obtiene reemplazando b por $-b$ en el corolario 4.9. Al generalizar a una combinación lineal de n variables aleatorias independientes, resulta el corolario 4.11.

Corolario 4.11: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes, entonces

$$\sigma^2_{a_1X_1+a_2X_2+\cdots+a_nX_n} = a_1^2\sigma^2_{X_1} + a_2^2\sigma^2_{X_2} + \cdots + a_n^2\sigma^2_{X_n}.$$

Ejemplo 4.22: Si X y Y son variables aleatorias con varianzas $\sigma^2_X = 2$ y $\sigma^2_Y = 4$ y covarianza $\sigma_{XY} = -2$, calcule la varianza de la variable aleatoria $Z = 3X - 4Y + 8$.

***Solución*:**

$$\begin{aligned}\sigma^2_Z &= \sigma^2_{3X-4Y+8} = \sigma^2_{3X-4Y} \quad \text{(por el corolario 4.6)}\\ &= 9\sigma^2_x + 16\sigma^2_y - 24\sigma_{xy} \quad \text{(por el teorema 4.9)}\\ &= (9)(2) + (16)(4) - (24)(-2) = 130.\end{aligned}$$

Ejemplo 4.23: Denotemos con X y Y la cantidad de dos tipos diferentes de impurezas en un lote de cierto producto químico. Suponga que X y Y son variables aleatorias independientes con varianzas $\sigma^2_X = 2$ y $\sigma^2_Y = 3$. Calcule la varianza de la variable aleatoria $Z = 3X - 2Y + 5$.

***Solución*:**

$$\begin{aligned}\sigma^2_Z &= \sigma^2_{3X-2Y+5} = \sigma^2_{3X-2Y} \quad \text{(por el corolario 4.6)}\\ &= 9\sigma^2_x + 4\sigma^2_y \quad \text{(por el corolario 4.10)}\\ &= (9)(2) + (4)(3) = 30.\end{aligned}$$

¿Qué sucede si la función es no lineal?

En las secciones anteriores estudiamos propiedades de funciones lineales de variables aleatorias por razones muy importantes. En los capítulos 8 a 15 se estudiarán y ejemplificarán problemas de la vida real, en los cuales el analista construye un **modelo lineal** para describir un conjunto de datos y, en consecuencia, describir o explicar el comportamiento de un fenómeno científico. Así que resulta natural que encontremos los valores esperados y las varianzas de combinaciones lineales de variables aleatorias. Sin embargo, hay situaciones en que las propiedades de las funciones **no lineales** de variables aleatorias se vuelven importantes. En efecto, hay muchos fenómenos científicos de naturaleza no lineal, donde el modelado estadístico que utiliza funciones no lineales adquiere gran importancia. De hecho, en el capítulo 12 se estudia el modelado de los que se han convertido en modelos estándar no lineales. En realidad, incluso una función simple de variables aleatorias, como $Z = X/Y$, ocurre con bastante frecuencia en la prác-

tica, y a diferencia del caso del valor esperado de las combinaciones lineales de variables aleatorias, no hay una simple regla general. Por ejemplo,

$$E(Z) = E(X/Y) \neq E(X)/E(Y),$$

excepto en circunstancias muy especiales.

El material dado por los teoremas 4.5 a 4.9 y los diversos corolarios son sumamente útiles, ya que no hay restricciones sobre la forma de la densidad o las funciones de probabilidad, aparte de la propiedad de independencia cuando ésta se requiere, como en los corolarios posteriores al teorema 4.9. Para ilustrar considere el ejemplo 4.23; la varianza de $Z = 3X - 2Y + 5$ no requiere restricciones en las distribuciones de las cantidades X y Y de los dos tipos de impurezas. Sólo se requiere la independencia entre X y Y. Por consiguiente, disponemos de la capacidad de calcular $\mu_{g(X)}$ y $\sigma^2_{g(X)}$ para cualquier función $g(\cdot)$ a partir de los principios iniciales establecidos en los teoremas 4.1 y 4.3, donde se supone que se **conoce** la distribución $f(x)$ correspondiente. Los ejercicios 4.40, 4.41 y 4.42, entre otros, ilustran el uso de tales teoremas. De modo que, si $g(x)$ es una función no lineal y se conoce la función de densidad (o función de probabilidad en el caso discreto), $\mu_{g(X)}$ y $\sigma^2_{g(X)}$ pueden evaluarse con exactitud. No obstante, como en el caso de las reglas dadas para combinaciones lineales, ¿habría reglas para funciones no lineales que se puedan utilizar cuando no se conoce la forma de la distribución de las variables aleatorias pertinentes?

En general, suponga que X es una variable aleatoria y que $Y = g(x)$. La solución general para $E(Y)$ o $\text{Var}(Y)$ puede ser difícil y depende de la complejidad de la función $g(\cdot)$. Sin embargo, hay aproximaciones diponibles que dependen de una aproximación lineal de la función $g(x)$. Por ejemplo, suponga que denotamos $E(X)$ como μ y $\text{Var}(X) = \sigma^2_X$. Entonces, una aproximación a las series de Taylor de $g(x)$ alrededor de $X = \mu_X$ da

$$g(x) = g(\mu_x) + \left.\frac{\partial g(x)}{\partial x}\right|_{x=\mu_X} (x - \mu_x) + \left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=\mu_X} \frac{(x-\mu_x)^2}{2} + \cdots.$$

Como resultado, si truncamos después el término lineal y tomamos el valor esperado de ambos lados, obtenemos $E[g(X)] \approx g(\mu_X)$, que ciertamente es intuitivo y en algunos casos ofrece una aproximación razonable. No obstante, si incluimos el término de segundo orden de la serie de Taylor, entonces tenemos un ajuste de segundo orden para esta *aproximación de primer orden* como sigue:

Aproximación de $E[g(X)]$

$$E[g(X)] \approx g(\mu_x) + \left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=\mu_X} \frac{\sigma^2_x}{2}.$$

Ejemplo 4.24: Dada la variable aleatoria X con media μ_X y varianza σ^2_X, determine la aproximación de segundo orden para $E(e^X)$.

***Solución*:** Como $\frac{\partial e^x}{\partial x} = e^x$ y $\frac{\partial^2 e^x}{\partial x^2} = e^x$, obtenemos $E(e^X) \approx e^{\mu_X}(1 + \sigma^2_x/2)$. ■

De manera similar, podemos desarrollar una aproximación para $\text{Var}[g(x)]$ tomando la varianza de ambos lados de la expansión de la serie de Taylor de primer orden de $g(x)$.

Aproximación de $\text{Var}[g(X)]$

$$\text{Var}[g(X)] \approx \left[\frac{\partial g(x)}{\partial x}\right]^2_{x=\mu_X} \sigma^2_x.$$

Ejemplo 4.25: Dada la variable aleatoria X, como en el ejemplo 4.24, determine una fórmula aproximada para $\text{Var}[g(x)]$.

***Solución*:** De nuevo, $\frac{\partial e^x}{\partial x} = e^x$ por lo tanto, $\text{Var}(X) \approx e^{2\mu_X}\,\sigma_x^2$. ■

Estas aproximaciones se pueden extender a las funciones no lineales de más de una variable aleatoria.

Dado un conjunto de variables aleatorias independientes $X_1, X_2, \ldots, X_k$ con medias μ_1, $\mu_2, \ldots, \mu_k$ y varianzas $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, respectivamente, sea

$$Y = h(X_1, X_2, \ldots, X_k)$$

una función no lineal; entonces tenemos las siguientes aproximaciones para $E(Y)$ y $\text{Var}(Y)$:

$$E(Y) \approx h(\mu_1, \mu_2, \ldots, \mu_k) + \sum_{i=1}^{k} \frac{\sigma_i^2}{2} \left[\frac{\partial^2 h(x_1, x_2, \ldots, x_k)}{\partial x_i^2}\right]\Bigg|_{x_i = \mu_i,\ 1 \le i \le k},$$

$$\text{Var}(Y) \approx \sum_{i=1}^{k} \left[\frac{\partial h(x_1, x_2, \ldots, x_k)}{\partial x_i}\right]^2\Bigg|_{x_i = \mu_i,\ 1 \le i \le k} \sigma_i^2.$$

Ejemplo 4.26: Considere dos variables aleatorias independientes X y Z, con medias μ_X, μ_Z y varianzas σ_X^2 y σ_Z^2, respectivamente. Considere una variable aleatoria

$$Y = X/Z.$$

Determine aproximaciones para $E(Y)$ y $\text{Var}(Y)$.

***Solución*:** Para $E(Y)$, debemos usar $\frac{\partial y}{\partial x} = \frac{1}{z}$ y $\frac{\partial y}{\partial z} = -\frac{x}{z^2}$. Por consiguiente,

$$\frac{\partial^2 y}{\partial x^2} = 0 \quad \text{y} \quad \frac{\partial^2 y}{\partial z^2} = \frac{2x}{z^3}.$$

Como resultado,

$$E(Y) \approx \frac{\mu_X}{\mu_Z} + \frac{\mu_X}{\mu_Z^3}\sigma_Z^2 = \frac{\mu_X}{\mu_Z}\left(1 + \frac{\sigma_Z^2}{\mu_Z^2}\right),$$

y la aproximación para la varianza de Y está dada por

$$\text{Var}(Y) \approx \frac{1}{\mu_Z^2}\sigma_X^2 + \frac{\mu_X^2}{\mu_Z^4}\sigma_Z^2 = \frac{1}{\mu_Z^2}\left(\sigma_X^2 + \frac{\mu_X^2}{\mu_Z^2}\sigma_Z^2\right).$$

■

4.4 Teorema de Chebyshev

En la sección 4.2 establecimos que la varianza de una variable aleatoria nos dice algo acerca de la variabilidad de las observaciones con respecto a la media. Si una variable aleatoria tiene una varianza o desviación estándar pequeña, esperaríamos que la mayoría de los valores se agrupen alrededor de la media. Por lo tanto, la probabilidad de que una variable aleatoria tome un valor dentro de cierto intervalo alrededor de la media es mayor que para una variable aleatoria similar con una desviación estándar mayor. Si pensamos en la probabilidad en términos de área, esperaríamos una distribución continua con un valor grande de σ para indicar una variabilidad mayor y, por lo tanto, esperaríamos que el área esté más extendida, como en la figura 4.2(a). Una distribución con una desviación estándar pequeña debería tener la mayor parte de su área cercana a μ, como en la figura 4.2(b).

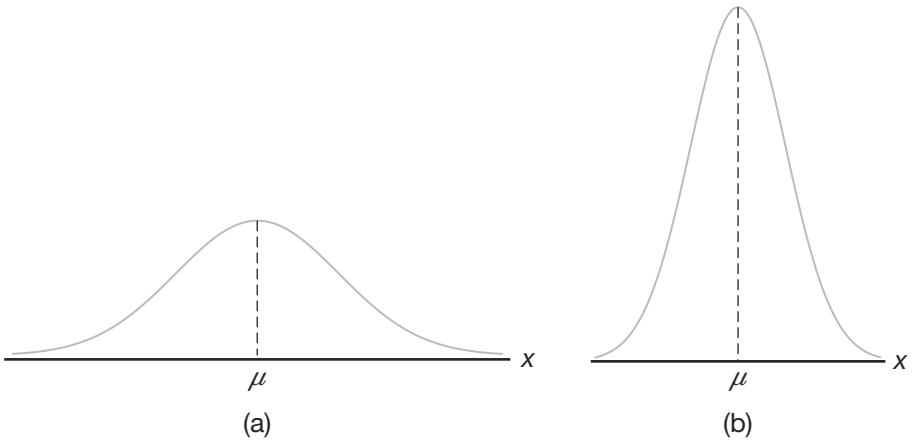


Figura 4.2: Variabilidad de observaciones continuas alrededor de la media.

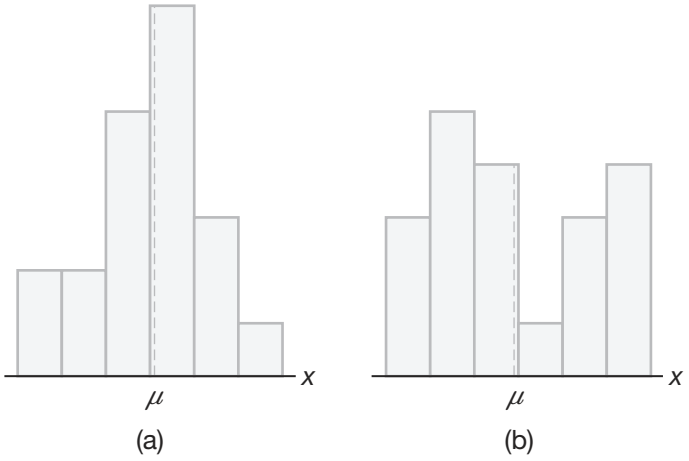


Figura 4.3: Variabilidad de observaciones discretas alrededor de la media.

Podemos argumentar lo mismo para una distribución discreta. En el histograma de probabilidad de la figura 4.3(b) el área se extiende mucho más que en la figura 4.3(a), lo cual indica una distribución más variable de mediciones o resultados.

El matemático ruso P. L. Chebyshev (1821-1894) descubrió que la fracción del área entre cualesquiera dos valores simétricos alrededor de la media está relacionada con la desviación estándar. Como el área bajo una curva de distribución de probabilidad, o la de un histograma de probabilidad, suma 1, el área entre cualesquiera dos números es la probabilidad de que la variable aleatoria tome un valor entre estos números.

El siguiente teorema, planteado por Chebyshev, ofrece una estimación conservadora de la probabilidad de que una variable aleatoria tome un valor dentro de k desviaciones estándar de su media para cualquier número real k.

Teorema 4.10: (**Teorema de Chebyshev**) La probabilidad de que cualquier variable aleatoria X tome un valor dentro de k desviaciones estándar de la media es de al menos $1 - 1/k^2$. Es decir,

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

Para $k = 2$ el teorema establece que la variable aleatoria X tiene una probabilidad de al menos $1-1/2^2 = 3/4$ de caer dentro de dos desviaciones estándar a partir de la media; es decir, que tres cuartas partes o más de las observaciones de cualquier distribución se localizan en el intervalo $\mu \pm 2\sigma$. De manera similar, el teorema afirma que al menos ocho novenos de las observaciones de cualquier distribución caen en el intervalo $\mu \pm 3\sigma$.

Ejemplo 4.27: Una variable aleatoria X tiene una media $\mu = 8$, una varianza $\sigma^2 = 9$ y una distribución de probabilidad desconocida. Calcule

a) $P(-4 < X < 20)$,

b) $P(|X - 8| \geq 6)$.

***Solución*:** *a*) $P(-4 < X < 20) = P[8 - (4)(3) < X < 8 + (4)(3)] \geq \frac{15}{16}$.

b) $P(|X - 8| \geq 6) = 1 - P(|X - 8| < 6) = 1 - P(-6 < X - 8 < 6)$

$$= 1 - P[8 - (2)(3) < X < 8 + (2)(3)] \leq \frac{1}{4}.$$

El teorema de Chebyshev tiene validez para cualquier distribución de observaciones, por lo cual los resultados generalmente son débiles. El valor que proporciona el teorema es sólo un límite inferior, es decir, sabemos que la probabilidad de una variable aleatoria que cae dentro de dos desviaciones estándar de la media *no puede ser menor* que 3/4, pero nunca sabemos cuánto podría ser en realidad. Sólo cuando conocemos la distribución de probabilidad podemos determinar probabilidades exactas. Por esta razón llamamos al teorema resultado de *distribución libre*. Cuando se supongan distribuciones específicas, como ocurrirá en los siguientes capítulos, los resultados serán menos conservadores. El uso del teorema de Chebyshev se restringe a situaciones donde se desconoce la forma de la distribución.

Ejercicios

4.53 Remítase al ejercicio 4.35 de la página 127 y calcule la media y la varianza de la variable aleatoria discreta $Z = 3X - 2$, donde X representa el número de errores por 100 líneas de código.

4.54 Use el teorema 4.5 y el corolario 4.6 para calcular la media y la varianza de la variable aleatoria $Z = 5X + 3$, donde X tiene la distribución de probabilidad del ejercicio 4.36 de la página 127.

4.55 Suponga que una tienda de abarrotes compra 5 envases de leche descremada al precio de mayoreo de \$1.20 por envase y la vende a \$1.65 por envase. Después de la fecha de caducidad, la leche que no se vende se retira de los anaqueles y el tendero recibe un crédito del distribuidor igual a tres cuartas partes del precio de mayoreo. Si la distribución de probabilidad de la variable aleatoria es X y el número de envases que se venden de este lote es

x	0	1	2	3	4	5
$f(x)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{3}{15}$

calcule la utilidad esperada.

4.56 Repita el ejercicio 4.43 de la página 127 aplicando el teorema 4.5 y el corolario 4.6.

4.57 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	−3	6	9
$f(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Calcule $E(X)$ y $E(X^2)$ y luego utilice estos valores para evaluar $E[(2X + 1)^2]$.

4.58 El tiempo total que una adolescente utiliza su secadora de pelo durante un año, medido en unidades de 100 horas, es una variable aleatoria continua X que tiene la siguiente función de densidad

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{en otro caso.} \end{cases}$$

Utilice el teorema 4.6 para evaluar la media de la variable aleatoria $Y = 60X^2 + 39X$, donde Y es igual al número de kilowatts-hora que gasta al año.

4.59 Si una variable aleatoria X se define de manera que

$$E[(X - 1)^2] = 10 \quad \text{y} \quad E[(X - 2)^2] = 6,$$

calcule μ y σ^2.

4.60 Suponga que X y Y son variables aleatorias independientes que tienen la siguiente distribución de probabilidad conjunta

		x	
$f(x, y)$		2	4
y	1	0.10	0.15
	3	0.20	0.30
	5	0.10	0.15

Calcule

a) $E(2X - 3Y)$;
b) $E(XY)$.

4.61 Use el teorema 4.7 para evaluar $E(2XY^2 - X^2Y)$ en la distribución de probabilidad conjunta que se muestra en la tabla 3.1 de la página 96.

4.62 Si X y Y son variables aleatorias independientes con varianzas $\sigma_X^2 = 5$ y $\sigma_Y^2 = 3$, calcule la varianza de la variable aleatoria $Z = -2X + 4Y - 3$.

4.63 Repita el ejercicio 4.62 si X y Y no son independientes y $\sigma_{XY} = 1$

4.64 Suponga que X y Y son variables aleatorias independientes con densidades de probabilidad y

$$g(x) = \begin{cases} \frac{8}{x^3}, & x > 2, \\ 0, & \text{en otro caso,} \end{cases}$$

y

$$h(y) = \begin{cases} 2y, & 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el valor esperado de $Z = XY$.

4.65 Sea X el número que resulta cuando se lanza un dado rojo y Y el número que resulta cuando se lanza un dado verde. Calcule

a) $E(X + Y)$;
b) $E(X - Y)$;
c) $E(XY)$.

4.66 Sea X el número que resulta cuando se lanza un dado verde y Y el número que resulta cuando se lanza un dado rojo. Calcule la varianza de la variable aleatoria

a) $2X - Y$;
b) $X + 3Y - 5$.

4.67 Si la función de densidad conjunta de X y Y está dada por

$$f(x, y) = \begin{cases} \frac{2}{7}(x + 2y), & 0 < x < 1, \ 1 < y < 2, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule el valor esperado de $g(X, Y) = \frac{X}{Y^3} + X^2 Y$.

4.68 Se sabe que la potencia P en watts que se disipa en un circuito eléctrico con resistencia R está dada por $P = I^2R$, donde I es la corriente en amperes y R es una constante fija en 50 ohms. Sin embargo, I es una variable aleatoria con $\mu_I = 15$ amperes y $\sigma_I^2 = 0.03$ amperes^2. Dé aproximaciones numéricas a la media y a la varianza de la potencia P.

4.69 Considere el ejercicio de repaso 3.77 de la página 108. Las variables aleatorias X y Y representan el número de vehículos que llegan a dos esquinas de calles separadas durante cierto periodo de 2 minutos en el día. La distribución conjunta es

$$f(x, y) = \left(\frac{1}{4^{(x+y)}}\right)\left(\frac{9}{16}\right),$$

para $x = 0, 1, 2,...$, y $y = 0, 1, 2,...$

a) Determine $E(X)$, $E(Y)$, Var(X) y Var(Y).
b) Considere que $Z = X + Y$ es la suma de ambas. Calcule $E(Z)$ y Var(Z).

4.70 Considere el ejercicio de repaso 3.64 de la página 107. Hay dos líneas de servicio. Las variables aleatorias X y Y son las proporciones del tiempo que la línea 1 y la línea 2 están en funcionamiento, respectivamente. La función de densidad de probabilidad conjunta para (X, Y) está dada por

$$f(x, y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 \le x, \ y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine si X y Y son independientes o no.

b) Se tiene interés por saber algo acerca de la proporción de $Z = X + Y$, la suma de las dos proporciones. Calcule $E(X + Y)$. También calcule $E(XY)$.
c) Calcule $\text{Var}(X)$, $\text{Var}(Y)$ y $\text{Cov}(X,Y)$.
d) Calcule $\text{Var}(X + Y)$.

4.71 El periodo Y en minutos que se requiere para generar un reflejo humano ante el gas lacrimógeno tiene la siguiente función de densidad

$$f(y) = \begin{cases} \frac{1}{4}e^{-y/4}, & 0 \le y < \infty, \\ 0, & \text{en otro caso.} \end{cases}$$

a) ¿Cuál es el tiempo medio para el reflejo?
b) Calcule $E(Y^2)$ y $\text{Var}(Y)$.

4.72 Una empresa industrial desarrolló una máquina de limpiar alfombras con buen rendimiento de combustible porque limpia más superficie de alfombra en menos tiempo. Se tiene interés por una variable aleatoria Y, la cantidad en galones por minuto que ofrece. Se sabe que la función de densidad está dada por

$$f(y) = \begin{cases} 1, & 7 \le y \le 8, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine la función de densidad.
b) Calcule $E(Y)$, $E(Y^2)$ y $\text{Var}(Y)$.

4.73 Para la situación del ejercicio 4.72 calcule $E(e^Y)$ utilizando el teorema 4.1, es decir, mediante el uso de

$$E(e^Y) = \int_7^8 e^y f(y)\, dy.$$

Luego, calcule $E(e^Y)$ sin utilizar $f(y)$. En su lugar utilice el ajuste de segundo orden para la aproximación de primer orden de $E(e^Y)$. Comente al respecto.

4.74 Considere nuevamente la situación del ejercicio 4.72, donde se le pide calcular $\text{Var}(e^Y)$. Utilice los teoremas 4.2 y 4.3 y defina $Z = e^Y$. En consecuencia, utilice las condiciones del ejercicio 4.73 para calcular

$$\text{Var}(Z) = E(Z^2) - [E(Z)]^2.$$

Luego hágalo sin utilizar $f(y)$. En su lugar utilice la aproximación de primer orden a las series de Taylor para $\text{Var}(e^Y)$. ¡Comente al respecto!

4.75 Una empresa eléctrica fabrica una bombilla de luz de 100 watts que, de acuerdo con las especificaciones escritas en la caja, tiene una vida media de 900 horas con una desviación estándar de 50 horas. A lo sumo, ¿qué porcentaje de las bombillas no duran al menos 700 horas? Suponga que la distribución es simétrica alrededor de la media.

4.76 En una planta de ensamble automotriz se crean 70 nuevos puestos de trabajo y se presentan 1000 aspirantes. Para seleccionar entre los aspirantes a los 70 mejores la armadora aplica un examen que abarca habilidad mecánica, destreza manual y capacidad matemática. La calificación media de este examen resulta ser 60 y las calificaciones tienen una desviación estándar de 6. ¿Una persona que obtiene una calificación de 84 puede obtener uno de los puestos? [*Sugerencia*: Utilice el teorema de Chebyshev]. Suponga que la distribución es simétrica alrededor de la media.

4.77 Una variable aleatoria X tiene una media $\mu = 10$ y una varianza $\sigma^2 = 4$. Utilice el teorema de Chebyshev para calcular

a) $P(|X - 10| \ge 3)$;
b) $P(|X - 10| < 3)$;
c) $P(5 < X < 15)$;
d) el valor de la constante c tal que

$$P(|X - 10| \ge c) \le 0.04.$$

4.78 Calcule $P(\mu - 2\sigma < X < \mu + 2\sigma)$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} 6x(1-x), & 0 < x < 1, \\ 0, & \text{en otro caso,} \end{cases}$$

y compare con el resultado dado por el teorema de Chebyshev.

Ejercicios de repaso

4.79 Demuestre el teorema de Chebyshev.

4.80 Calcule la covarianza de las variables aleatorias X y Y que tienen la siguiente función de densidad de probabilidad conjunta

$$f(x, y) = \begin{cases} x + y, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

4.81 Remítase a las variables aleatorias cuya función de densidad de probabilidad conjunta está dada en el ejercicio 3.47 de la página 105 y calcule la cantidad promedio de queroseno que queda en el tanque al final del día.

4.82 Suponga que la duración X en minutos de un tipo específico de conversación telefónica es una variable aleatoria con función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{1}{5}e^{-x/5}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Determine la duración media $E(X)$ de este tipo de conversación telefónica.
b) Calcule la varianza y la desviación estándar de X.
c) Calcule $E[(X + 5)^2]$.

4.83 Remítase a las variables aleatorias cuya función de densidad conjunta está dada en el ejercicio 3.41 de la página 105 y calcule la covarianza entre el peso de las cremas y el peso de los chiclosos en estas cajas de chocolates.

4.84 Remítase a las variables aleatorias cuya función de densidad de probabilidad conjunta está dada en el ejercicio 3.41 de la página 105 y calcule el peso esperado para la suma de las cremas y los chiclosos si uno compra una caja de tales chocolates.

4.85 Suponga que se sabe que la vida de un compresor particular X, en horas, tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{1}{900}e^{-x/900}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule la vida media del compresor.
b) Calcule $E(X^2)$.
c) Calcule la varianza y la desviación estándar de la variable aleatoria X.

4.86 Remítase a las variables aleatorias cuya función de densidad conjunta está dada en el ejercicio 3.40 de la página 105,

a) calcule μ_X y μ_Y;
b) calcule $E[(X + Y)/2]$.

4.87 Demuestre que $\text{Cov}(aX, bY) = ab\,\text{Cov}(X, Y)$.

4.88 Considere la función de densidad del ejercicio de repaso 4.85. Demuestre que el teorema de Chebyshev es válido para $k = 2$ y $k = 3$.

4.89 Considere la siguiente función de densidad conjunta

$$f(x, y) = \begin{cases} \frac{16y}{x^3}, & x > 2,\ 0 < y < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule el coeficiente de correlación ρ_{XY}.

4.90 Considere las variables aleatorias X y Y del ejercicio 4.63 de la página 138. Calcule ρ_{XY}.

4.91 La utilidad de un distribuidor, en unidades de \$5000, por un automóvil nuevo es una variable aleatoria X que tiene la siguiente función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 \le x \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule la varianza de la utilidad del distribuidor.
b) Demuestre que el teorema de Chebyshev es válido para $k = 2$ con la función de densidad anterior.
c) ¿Cuál es la probabilidad de que la utilidad exceda \$500?

4.92 Considere el ejercicio 4.10 de la página 117. ¿Se puede decir que las calificaciones dadas por los dos expertos son independientes? Explique su respuesta.

4.93 Los departamentos de marketing y de contabilidad de una empresa determinaron que si la empresa comercializa su producto creado recientemente, su contribución a las utilidades de la empresa durante los próximos 6 meses será la siguiente:

Contribución a las utilidades	Probabilidad
−\$5,000	0.2
\$10,000	0.5
\$30,000	0.3

¿Cuál es la utilidad esperada de la empresa?

4.94 En un sistema de apoyo para el programa espacial estadounidense, un componente crucial único funciona sólo 85 por ciento del tiempo. Para aumentar la confiabilidad del sistema se decidió instalar tres componentes paralelos, de manera que el sistema falle sólo si todos fallan. Suponga que los componentes actúan de forma independiente y que son equivalentes en el sentido de que 3 de ellos tienen una tasa de éxito de 85 por ciento. Considere la variable aleatoria X como el número de componentes de cada tres que fallan.

a) Escriba una función de probabilidad para la variable aleatoria X.
b) ¿Cuál es $E(X)$ (es decir, el número medio de componentes de cada tres que fallan)?
c) ¿Cuál es $\text{Var}(X)$?
d) ¿Cuál es la probabilidad de que el sistema completo sea exitoso?
e) ¿Cuál es la probabilidad de que falle el sistema?
f) Si se desea que el sistema tenga una probabilidad de éxito de 0.99, ¿son suficientes los tres componentes? Si no lo son, ¿cuántos se requerirían?

4.95 En los negocios es importante planear y llevar a cabo investigación para anticipar lo que ocurrirá al final del año. La investigación sugiere que el espectro de utilidades (pérdidas) de cierta empresa, con sus respectivas probabilidades, es el siguiente:

Utilidad	Probabilidad
−\$15,000	0.05
\$0	0.15
\$15,000	0.15
\$25,000	0.30
\$40,000	0.15
\$50,000	0.10
\$100,000	0.05
\$150,000	0.03
\$200,000	0.02

a) ¿Cuál es la utilidad esperada?
b) Determine la desviación estándar de las utilidades.

4.96 Mediante un conjunto de datos, y por la amplia investigación, se sabe que la cantidad de tiempo que cierto empleado de una empresa llega tarde a trabajar, medido en segundos, es una variable aleatoria X con la siguiente función de densidad

$$f(x) = \begin{cases} \frac{3}{(4)(50^3)}(50^2 - x^2), & -50 \le x \le 50, \\ 0, & \text{en otro caso.} \end{cases}$$

En otras palabras, él no sólo llega ligeramente retrasado a veces, sino que también puede llegar temprano a trabajar.
a) Calcule el valor esperado del tiempo en segundos que llega tarde.
b) Calcule $E(X^2)$.
c) ¿Cuál es la desviación estándar del tiempo en que llega tarde?

4.97 Un camión de carga viaja desde el punto A hasta el punto B y regresa por la misma ruta diariamente. Hay cuatro semáforos en la ruta. Sea X_1 el número de semáforos en rojo que el camión encuentra cuando va de A a B y X_2 el número de los que encuentra en el viaje de regreso. Los datos recabados durante un periodo largo sugieren que la distribución de probabilidad conjunta para (X_1, X_2) está dada por

	x_2				
x_1	0	1	2	3	4
0	0.01	0.01	0.03	0.07	0.01
1	0.03	0.05	0.08	0.03	0.02
2	0.03	0.11	0.15	0.01	0.01
3	0.02	0.07	0.10	0.03	0.01
4	0.01	0.06	0.03	0.01	0.01

a) Determine la densidad marginal de X_1.
b) Determine la densidad marginal de X_2.
c) Determine la distribución de densidad condicional de X_1 dado que $X_2 = 3$.
d) Determine $E(X_1)$.
e) Determine $E(X_2)$.
f) Determine $E(X_1 | X_2 = 3)$.
g) Determine la desviación estándar de X_1.

4.98 Una tienda de abarrotes tiene dos sitios separados en sus instalaciones donde los clientes pueden pagar cuando se marchan. Estos dos lugares tienen dos cajas registradoras y dos empleados que atienden a los clientes que van a pagar. Sea X el número de la caja registradora que se utiliza en un momento específico en el sitio 1 y Y el número de la caja registradora que se utiliza en el mismo momento en el sitio 2. La función de probabilidad conjunta está dada por

	y		
x	0	1	2
0	0.12	0.04	0.04
1	0.08	0.19	0.05
2	0.06	0.12	0.30

a) Determine la densidad marginal de X y de Y, así como la distribución de probabilidad de X, dado que $Y = 2$.
b) Determine $E(X)$ y $\text{Var}(X)$.
c) Determine $E(X | Y = 2)$ y $\text{Var}(X | Y = 2)$.

4.99 Considere un transbordador que puede llevar tanto autobuses como automóviles en un recorrido a través de una vía fluvial. Cada viaje cuesta al propietario aproximadamente \$10. La tarifa por automóvil es de \$3 y por autobús es de \$8. Sean X y Y el número de autobuses y automóviles, respectivamente, que se transportan en un viaje específico. La distribución conjunta de X y Y está dada por

	x		
y	0	1	2
0	0.01	0.01	0.03
1	0.03	0.08	0.07
2	0.03	0.06	0.06
3	0.07	0.07	0.13
4	0.12	0.04	0.03
5	0.08	0.06	0.02

Calcule la utilidad esperada para el viaje del transbordador.

4.100 Como veremos en el capítulo 12, los métodos estadísticos asociados con los modelos lineal y no lineal son muy importantes. De hecho, a menudo las funciones exponenciales se utilizan en una amplia gama de problemas científicos y de ingeniería. Considere un modelo que se ajusta a un conjunto de datos que implica los valores medidos k_1 y k_2, y una respuesta específica Y a las mediciones. El modelo postulado es

$$\hat{Y} = e^{b_0 + b_1 k_1 + b_2 k_2},$$

donde $\hat{Y}$ denota el **valor estimado de *Y***, k_1 y k_2 son valores fijos y b_0, b_1 y b_2 son **estimados** de constantes y, por lo tanto, variables aleatorias. Suponga que tales variables aleatorias son independientes y use la fórmula aproximada para la varianza de una función no lineal de más de una variable. Dé una expresión para $\text{Var}(\hat{Y})$. Suponga que se conocen las medias de b_0, b_1 y b_2 y que son β_0, β_1 y β_2, y también suponga que se conocen las varianzas de b_0, b_1 y b_2 y que son σ_0^2, σ_1^2 y σ_2^2.

4.101 Considere el ejercicio de repaso 3.73 de la página 108, el cual implica Y, la proporción de impurezas en un lote, donde la función de densidad está dada por

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Calcule el porcentaje esperado de impurezas.
b) Calcule el valor esperado de la proporción de la calidad del material (es decir, calcule $E(1 - Y)$).
c) Calcule la varianza de la variable aleatoria $Z = 1 - Y$.

4.102 **Proyecto:** Sea $X =$ número de horas que cada estudiante del grupo durmió la noche anterior. Cree una variable discreta utilizando los siguientes intervalos arbitrarios:

$X < 3$, $3 \le X < 6$, $6 \le X < 9$ y $X \ge 9$.

a) Estime la distribución de probabilidad para X.
b) Calcule la media estimada y la varianza para X.

4.5 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

El material que se cubrió en este capítulo es fundamental, como el contenido del capítulo 3. Mientras que en el capítulo 3 nos concentramos en las características generales de una distribución de probabilidad, en el presente capítulo definimos cantidades importantes o *parámetros* que caracterizan la naturaleza general del sistema. La **media** de una distribución refleja una *tendencia central*, en tanto que la **varianza** o la **desviación estándar** reflejan *variabilidad* en el sistema. Además, la covarianza refleja la tendencia de dos variables aleatorias a "moverse juntas" en un sistema. Estos importantes parámetros serán fundamentales en el estudio de los siguientes capítulos.

El lector debería comprender que el tipo de distribución a menudo está determinado por el contexto científico. Sin embargo, los valores del parámetro necesitan estimarse a partir de datos científicos. Por ejemplo, en el caso del ejercicio de repaso 4.85 el fabricante del compresor podría saber (material que se presentará en el capítulo 6), por su experiencia y conocimiento del tipo de compresor, que la naturaleza de la distribución es como se indica en el ejercicio. Pero la media $\mu = 900$ **se estimaría** a partir de la experimentación con la máquina. Aunque aquí se da por conocido el valor del parámetro de 900, en situaciones reales eso no ocurrirá sin el uso de datos experimentales. El capítulo 9 se dedica a la **estimación**.

Capítulo 5

Algunas distribuciones de probabilidad discreta

5.1 Introducción y motivación

La distribución de probabilidad discreta describe el comportamiento de una variable aleatoria, independientemente de si se representa de forma gráfica o mediante un histograma, en forma tabular o con una fórmula. A menudo las observaciones que se generan mediante diferentes experimentos estadísticos tienen el mismo tipo general de comportamiento. En consecuencia, las variables aleatorias discretas asociadas con estos experimentos se pueden describir esencialmente con la misma distribución de probabilidad y, por lo tanto, es posible representarlas usando una sola fórmula. De hecho, se necesitan sólo unas cuantas distribuciones de probabilidad importantes para describir muchas de las variables aleatorias discretas que se encuentran en la práctica.

Este conjunto de distribuciones en realidad describe varios fenómenos aleatorios de la vida real. Por ejemplo, en un estudio en el que se probó la eficacia de un nuevo fármaco, de todos los pacientes que lo utilizaron, el número de pacientes que se curaron se aproximó a una distribución binomial (sección 5.2). En un ejemplo en una industria, cuando se prueba una muestra de artículos seleccionados de un lote de producción, el número de productos defectuosos en la muestra por lo general se puede representar como una variable aleatoria hipergeométrica (sección 5.3). En un problema estadístico de control de calidad el experimentador señalará un cambio en la media del proceso cuando los datos observacionales excedan ciertos límites. El número de muestras requeridas para generar una falsa alarma sigue una distribución geométrica, que es un caso especial de distribución binomial negativa (sección 5.4). Por otro lado, el número de leucocitos de una cantidad fija de una muestra de la sangre de un individuo suele ser aleatorio y podría describirse mediante una distribución de Poisson (sección 5.5). En este capítulo se presentarán esas distribuciones de uso común con varios ejemplos.

5.2 Distribuciones binomial y multinomial

Con frecuencia un experimento consta de pruebas repetidas, cada una con dos resultados posibles que se pueden denominar **éxito** o **fracaso**. La aplicación más evidente tiene que ver con la prueba de artículos a medida que salen de una línea de ensamble, donde cada

prueba o experimento puede indicar si un artículo está o no defectuoso. Podemos elegir definir cualquiera de los resultados como éxito. El proceso se conoce como **proceso de Bernoulli** y cada ensayo se denomina **experimento de Bernoulli**. Por ejemplo, si extraemos cartas de una baraja y éstas no se reemplazan, cambian las probabilidades en la repetición de cada ensayo; es decir, la probabilidad de seleccionar una carta de corazones en la primera extracción es 1/4, pero en la segunda es una probabilidad condicional que tiene un valor de 13/51 o 12/51, dependiendo de si resulta un corazón en la primera extracción; entonces éste ya no sería considerado un conjunto de experimentos de Bernoulli.

El proceso de Bernoulli

En términos estrictos el proceso de Bernoulli se caracteriza por lo siguiente:

1. El experimento consta de ensayos repetidos.
2. Cada ensayo produce un resultado que se puede clasificar como éxito o fracaso.
3. La probabilidad de un éxito, que se denota con p, permanece constante de un ensayo a otro.
4. Los ensayos repetidos son independientes.

Considere el conjunto de experimentos de Bernoulli en el que se seleccionan tres artículos al azar de un proceso de producción, luego se inspeccionan y se clasifican como defectuosos o no defectuosos. Un artículo defectuoso se designa como un éxito. El número de éxitos es una variable aleatoria X que toma valores integrales de cero a 3. Los ocho resultados posibles y los valores correspondientes de X son

Resultado	*NNN*	*NDN*	*NND*	*DNN*	*NDD*	*DND*	*DDN*	*DDD*
x	0	1	1	1	2	2	2	3

Como los artículos se seleccionan de forma independiente y se asume que el proceso produce 25% de artículos defectuosos,

$$P(NDN) = P(N)P(D)P(N) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{64}.$$

Cálculos similares dan las probabilidades para los otros resultados posibles. La distribución de probabilidad de X es, por lo tanto,

x	0	1	2	3
$f(x)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

Distribución binomial

El número X de éxitos en n experimentos de Bernoulli se denomina **variable aleatoria binomial**. La distribución de probabilidad de esta variable aleatoria discreta se llama **distribución binomial** y sus valores se denotarán como $b(x; n, p)$, ya que dependen del número de ensayos y de la probabilidad de éxito en un ensayo dado. Por consiguiente, para la distribución de probabilidad de X el número de productos defectuosos es

$$P(X = 2) = f(2) = b\left(2; 3, \frac{1}{4}\right) = \frac{9}{64}.$$

Generalicemos ahora la ilustración anterior con el fin de obtener una fórmula para $b(x; n, p)$. Esto significa que deseamos encontrar una fórmula que dé la probabilidad de x éxitos en n ensayos para un experimento binomial. Empiece por considerar la probabilidad de x éxitos y $n - x$ fracasos en un orden específico. Como los ensayos son independientes, podemos multiplicar todas las probabilidades que corresponden a los diferentes resultados. Cada éxito ocurre con probabilidad p y cada fracaso con probabilidad $q = 1 - p$. Por lo tanto, la probabilidad para el orden específico es $p^x q^{n-x}$. Ahora debemos determinar el número total de puntos muestrales en el experimento que tienen x éxitos y $n - x$ fracasos. Este número es igual al número de particiones de n resultados en dos grupos con x en un grupo y $n - x$ en el otro, y se escribe $\binom{n}{x}$ como se presentó en la sección 2.3. Como estas particiones son mutuamente excluyentes, sumamos las probabilidades de todas las diferentes particiones para obtener la fórmula general o simplemente multiplicamos $p^x q^{n-x}$ por $\binom{n}{x}$.

Distribución binomial

Un experimiento de Bernoulli puede tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$. Entonces, la distribución de probabilidad de la variable aleatoria binomial X, el número de éxitos en n ensayos independientes, es

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Observe que cuando $n = 3$ y $p = 1/4$, la distribución de probabilidad de X, el número de artículos defectuosos, se escribe como

$$b\left(x; 3, \frac{1}{4}\right) = \binom{3}{x}\left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3,$$

en vez de la forma tabular de la página 144.

Ejemplo 5.1: La probabilidad de que cierta clase de componente sobreviva a una prueba de choque es de 3/4. Calcule la probabilidad de que sobrevivan exactamente 2 de los siguientes 4 componentes que se prueben.

***Solución*:** Si suponemos que las pruebas son independientes y $p = 3/4$ para cada una de las 4 pruebas, obtenemos

$$b\left(2; 4, \frac{3}{4}\right) = \binom{4}{2}\left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2 = \left(\frac{4!}{2!\,2!}\right)\left(\frac{3^2}{4^4}\right) = \frac{27}{128}.$$

¿De dónde proviene el nombre binomial?

La distribución binomial deriva su nombre del hecho de que los $n + 1$ términos en la expansión binomial de $(q + p)^n$ corresponden a los diversos valores de $b(x; n, p)$ para $x = 0, 1, 2, \ldots, n$. Es decir,

$$\begin{aligned}(q+p)^n &= \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n \\ &= b(0; n, p) + b(1; n, p) + b(2; n, p) + \cdots + b(n; n, p).\end{aligned}$$

Dado que $p + q = 1$, vemos que

$$\sum_{x=0}^{n} b(x; n, p) = 1,$$

una condición que se debe cumplir para cualquier distribución de probabilidad.

Con frecuencia nos interesamos en problemas donde se necesita obtener $P(X < r)$ o $P(a \leq X \leq b)$. Las sumatorias binomiales

$$B(r; n, p) = \sum_{x=0}^{r} b(x; n, p)$$

se presentan en la tabla A.1 del apéndice para $n = 1, 2, ..., 20$, para valores seleccionados de p entre 0.1 y 0.9. Ilustramos el uso de la tabla A.1 con el siguiente ejemplo.

Ejemplo 5.2: La probabilidad de que un paciente se recupere de una rara enfermedad sanguínea es de 0.4. Si se sabe que 15 personas contrajeron la enfermedad, ¿cuál es la probabilidad de que *a*) sobrevivan al menos 10, *b*) sobrevivan de 3 a 8, y *c*) sobrevivan exactamente 5?

***Solución*:** Sea X el número de personas que sobreviven.

$$a) \quad P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662$$
$$= 0.0338$$

$$b) \quad P(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4)$$
$$= 0.9050 - 0.0271 = 0.8779$$

$$c) \quad P(X = 5) = b(5; 15, 0.4) = \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4)$$
$$= 0.4032 - 0.2173 = 0.1859$$

Ejemplo 5.3: Una cadena grande de tiendas al detalle le compra cierto tipo de dispositivo electrónico a un fabricante, el cual le indica que la tasa de dispositivos defectuosos es de 3%.

a) El inspector de la cadena elige 20 artículos al azar de un cargamento. ¿Cuál es la probabilidad de que haya al menos un artículo defectuoso entre estos 20?

b) Suponga que el detallista recibe 10 cargamentos en un mes y que el inspector prueba aleatoriamente 20 dispositivos por cargamento. ¿Cuál es la probabilidad de que haya exactamente tres cargamentos que contengan al menos un dispositivo defectuoso de entre los 20 seleccionados y probados?

***Solución*:** *a*) Denote con X el número de dispositivos defectuosos de los 20. Entonces X sigue una distribución $b(x; 20, 0.03)$. Por consiguiente,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - b(0; 20, 0.03)$$
$$= 1 - (0.03)^0(1 - 0.03)^{20-0} = 0.4562.$$

b) En este caso cada cargamento puede o no contener al menos un artículo defectuoso. Por lo tanto, el hecho de probar el resultado de cada cargamento puede considerarse como un experimento de Bernoulli con $p = 0.4562$ del inciso *a*). Si suponemos la independencia de un cargamento a otro, y si se denotamos con Y el número de cargamentos que contienen al menos un artículo defectuoso, Y sigue otra distribución bi-

nomial $b(y; 10, 0.4562)$. Por lo tanto,

$$P(Y = 3) = \binom{10}{3} 0.4562^3 (1 - 0.4562)^7 = 0.1602.$$

Áreas de aplicación

A partir de los ejemplos 5.1 a 5.3 debería quedar claro que la distribución binomial tiene aplicaciones en muchos campos científicos. Un ingeniero industrial está muy interesado en "la proporción de artículos defectuosos" en cierto proceso industrial. A menudo las medidas de control de calidad y los esquemas de muestreo para procesos se basan en la distribución binomial, la cual se aplica en cualquier situación industrial donde el resultado de un proceso es dicotómico y los resultados del proceso son independientes, y además la probabilidad de éxito se mantiene constante de una prueba a otra. La distribución binomial también se utiliza mucho en aplicaciones médicas y militares. En ambos casos un resultado de éxito o de fracaso es importante. Por ejemplo, la importancia del trabajo farmacéutico radica en poder determinar si un determinado fármaco "cura" o "no cura"; mientras que si se está probando la eficacia al lanzar un proyectil el resultado se interpretaría como "dar en el blanco" o "fallar".

Como la distribución de probabilidad de cualquier variable aleatoria binomial depende sólo de los valores que toman los parámetros n, p y q, parecería razonable suponer que la media y la varianza de una variable aleatoria binomial también dependen de los valores que toman tales parámetros. En realidad esto es cierto, y en la demostración del teorema 5.1 derivamos fórmulas generales que se pueden utilizar para calcular la media y la varianza de cualquier variable aleatoria binomial como funciones de n, p y q.

Teorema 5.1: La media y la varianza de la distribución binomial $b(x; n, p)$ son

$$\mu = np \text{ y } \sigma^2 = npq.$$

***Prueba*:** Representemos el resultado de la j-ésima prueba mediante una variable aleatoria de Bernoulli I_j, que toma los valores 0 y 1 con probabilidades q y p, respectivamente. Por lo tanto, en un experimento binomial el número de éxitos se escribe como la suma de las n variables indicadoras independientes. De aquí,

$$X = I_1 + I_2 + \cdots + I_n.$$

La media de cualquier I_j es $E(I_j) = (0)(q) + (1)(p) = p$. Por lo tanto, usando el corolario 4.4 de la página 131, la media de la distribución binomial es

$$\mu = E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = \underbrace{p + p + \cdots + p}_{n \text{ términos}} = np.$$

La varianza de cualquier I_j es $\sigma^2_{I_j} = E(I_j^2) - p^2 = (0)^2(q) + (1)^2(p) - p^2 = p(1-p) = pq$. Al ampliar el corolario 4.11 al caso de n variables de Bernoulli independientes, la varianza de la distribución binomial resulta como

$$\sigma^2_X = \sigma^2_{I_1} + \sigma^2_{I_2} + \cdots + \sigma^2_{I_n} = \underbrace{pq + pq + \cdots + pq}_{n \text{ términos}} = npq.$$

Ejemplo 5.4: Se conjetura que hay impurezas en 30% del total de pozos de agua potable de cierta comunidad rural. Para obtener información sobre la verdadera magnitud del problema se determina que debe realizarse algún tipo de prueba. Como es muy costoso probar todos los pozos del área, se eligen 10 al azar para someterlos a la prueba.

a) Si se utiliza la distribución binomial, ¿cuál es la probabilidad de que exactamente 3 pozos tengan impurezas, considerando que la conjetura es correcta?

b) ¿Cuál es la probabilidad de que más de 3 pozos tengan impurezas?

***Solución*:** *a*) Requerimos

$$b(3; 10, 0.3) = \sum_{x=0}^{3} b(x; 10, 0.3) - \sum_{x=0}^{2} b(x; 10, 0.3) = 0.6496 - 0.3828 = 0.2668.$$

b) En este caso $P(X > 3) = 1 - 0.6496 = 0.3504$. ∎

Ejemplo 5.5: Calcule la media y la varianza de la variable aleatoria binomial del ejemplo 5.2 y después utilice el teorema de Chebyshev (de la página 137) para interpretar el intervalo $\mu \pm 2\sigma$.

***Solución*:** Como el ejemplo 5.2 fue un experimento binomial con $n = 15$ y $p = 0.4$, por el teorema 5.1 tenemos

$$\mu = (15)(0.4) = 6 \text{ y } \sigma^2 = (15)(0.4)(0.6) = 3.6.$$

Al tomar la raíz cuadrada de 3.6 encontramos que $\sigma = 1.897$. Por lo tanto, el intervalo que se requiere es $6 \pm (2)(1.897)$, o de 2.206 a 9.794. El teorema de Chebyshev establece que el número de pacientes recuperados, de un total de 15 que contrajeron la enfermedad, tiene una probabilidad de al menos 3/4 de caer entre 2.206 y 9.794 o, como los datos son discretos, incluso entre 2 y 10. ∎

Hay soluciones en las que el cálculo de las probabilidades binomiales nos permitirían hacer inferencias científicas acerca de una población después de que se recaban los datos. El siguiente ejemplo es una ilustración de esto.

Ejemplo 5.6: Considere la situación del ejemplo 5.4. La idea de que el 30% de los pozos tienen impurezas es sólo una conjetura del consejo local del agua. Suponga que se eligen 10 pozos de forma aleatoria y resulta que 6 contienen impurezas. ¿Qué implica esto respecto de la conjetura? Utilice un enunciado de probabilidad.

***Solución*:** Primero debemos preguntar: "Si la conjetura es correcta, ¿podríamos haber encontrado 6 o más pozos con impurezas?"

$$P(X \geq 6) = \sum_{x=0}^{10} b(x; 10, 0.3) - \sum_{x=0}^{5} b(x; 10, 0.3) = 1 - 0.9527 = 0.0473.$$

En consecuencia, es poco probable (4.7% de probabilidad) que se encontrara que 6 o más pozos contenían impurezas si sólo 30% de ellos las contienen. Esto pone seriamente en duda la conjetura y sugiere que el problema de la impureza es mucho más grave. ∎

Como podrá darse cuenta el lector ahora, en muchas aplicaciones hay más de dos resultados posibles. Por ejemplo, en el campo de la genética el color de las crías de conejillos de Indias puede ser rojo, negro o blanco. Con frecuencia la dicotomía de "defectuoso" y "sin defectos" en casos de ingeniería es en realidad un simplificación excesiva. De hecho, a menudo hay más de dos categorías que caracterizan los artículos o las partes que salen de una línea de producción.

Experimentos multinomiales y la distribución multinomial

El experimento binomial se convierte en un **experimento multinomial** si cada prueba tiene más de dos resultados posibles. La clasificación de un producto fabricado como ligero, pesado o aceptable, y el registro de los accidentes en cierto crucero de acuerdo con el día de la semana, constituyen experimentos multinomiales. Extraer *con reemplazo* una carta de una baraja también es un experimento multinomial si los 4 palos son los resultados de interés.

En general, si un ensayo dado puede tener como consecuencia cualquiera de los k resultados posibles $E_1, E_2, ..., E_k$ con probabilidades $p_1, p_2, ..., p_k$, la **distribución multinomial** dará la probabilidad de que E_1 ocurra x_1 veces, E_2 ocurra x_2 veces... y E_k ocurra x_k veces en n ensayos independientes, donde

$$x_1 + x_2 + \cdots + x_k = n.$$

Denotaremos esta distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n).$$

Salta a la vista que $p_1 + p_2 + \cdots + p_k = 1$, pues el resultado de cada ensayo debe ser uno de los k resultados posibles.

Para derivar la fórmula general procedemos como en el caso binomial. Puesto que los ensayos son independientes, cualquier orden especificado que produzca x_1 resultados para E_1, x_2 para E_2,..., x_k para E_k ocurrirá con probabilidad $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$. El número total de ordenamientos que producen resultados similares para los n ensayos es igual al número de particiones de n artículos en k grupos con x_1 en el primer grupo, x_2 en el segundo grupo,..., y x_k en el k-ésimo grupo. Esto se puede hacer en

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}$$

formas. Como todas las particiones son mutuamente excluyentes y tienen la misma probabilidad de ocurrir, obtenemos la distribución multinomial multiplicando la probabilidad para un orden específico por el número total de particiones.

Distribución multinomial

Si un ensayo dado puede producir los k resultados $E_1, E_2, ..., E_k$ con probabilidades $p_1, p_2, \ldots, p_k$, entonces la distribución de probabilidad de las variables aleatorias $X_1, X_2, ..., X_k$, que representa el número de ocurrencias para $E_1, E_2, ..., E_k$ en n ensayos independientes, es

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

con

$$\sum_{i=1}^{k} x_i = n \text{ y } \sum_{i=1}^{k} p_i = 1.$$

La distribución multinomial deriva su nombre del hecho de que los términos de la expansión multinomial de $(p_1 + p_2 + ... + p_k)^n$ corresponden a todos los posibles valores de $f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n)$.

Ejemplo 5.7: La complejidad de las llegadas y las salidas de los aviones en un aeropuerto es tal que a menudo se utiliza la simulación por computadora para modelar las condiciones "ideales". Para un aeropuerto específico que tiene tres pistas se sabe que, en el escenario ideal, las probabilidades de que las pistas individuales sean utilizadas por un avión comercial que llega aleatoriamente son las siguientes:

Pista 1: $p_1 = 2/9$
Pista 2: $p_2 = 1/6$
Pista 3: $p_3 = 11/18$

¿Cuál es la probabilidad de que 6 aviones que llegan al azar se distribuyan de la siguiente manera?

Pista 1: 2 aviones
Pista 2: 1 avión
Pista 3: 3 aviones

***Solución*:** Si usamos la distribución multinomial, tenemos

$$f\left(2, 1, 3; \frac{2}{9}, \frac{1}{6}, \frac{11}{18}, 6\right) = \binom{6}{2, 1, 3}\left(\frac{2}{9}\right)^2\left(\frac{1}{6}\right)^1\left(\frac{11}{18}\right)^3$$

$$= \frac{6!}{2!1!3!} \cdot \frac{2^2}{9^2} \cdot \frac{1}{6} \cdot \frac{11^3}{18^3} = 0.1127.$$

Ejercicios

5.1 Una variable aleatoria X que toma los valores x_1, x_2,..., x_k se denomina variable aleatoria discreta uniforme si su función de masa de probabilidad es $f(x) = \frac{1}{k}$ para todas las variables x_1, x_2,…, x_k y 0 en cualquier otro caso. Calcule la media y la varianza de X.

5.2 Se entregan dos altavoces idénticos a 12 personas y se les pide que los escuchen para determinar si hay alguna diferencia entre ellos. Suponga que sus respuestas son simplemente conjeturas. Calcule la probabilidad de que tres personas afirmen haber detectado una diferencia entre los dos altavoces.

5.3 De un equipo de 10 empleados, y mediante la selección al azar de una etiqueta contenida en una caja que contiene 10 etiquetas numeradas del 1 al 10, se elige a uno para que supervise cierto proyecto. Calcule la fórmula para la distribución de probabilidad de X que represente el número en la etiqueta que se saca. ¿Cuál es la probabilidad de que el número que se extrae sea menor que 4?

5.4 En cierto distrito de la ciudad se establece que la causa de 75% de todos los robos es la necesidad de dinero para comprar drogas. Calcule la probabilidad de que entre los siguientes cinco casos de robo que se reporten en este distrito,

a) exactamente 2 sean resultado de la necesidad de dinero para comprar drogas;
b) a lo sumo 3 resulten de la necesidad de dinero para comprar drogas.

5.5 De acuerdo con *Chemical Engineering Progress* (noviembre de 1990), aproximadamente 30% de todas las fallas de operación en las tuberías de plantas químicas son ocasionadas por errores del operador.

a) ¿Cuál es la probabilidad de que de las siguientes 20 fallas en las tuberías al menos 10 se deban a un error del operador?
b) ¿Cuál es la probabilidad de que no más de 4 de 20 fallas se deban a un error del operador?
c) Suponga que, para una planta específica, de la muestra aleatoria de 20 de tales fallas exactamente 5 son errores de operación. ¿Considera que la cifra de 30% anterior se aplique a esta planta? Comente su respuesta.

5.6 De acuerdo con una encuesta de la *Administrative Management Society*, la mitad de las empresas estadounidenses da a sus empleados 4 semanas de vacaciones después de 15 años de servicio en la empresa. Calcule la probabilidad de que, de 6 empresas encuestadas al azar, el número que da a sus empleados 4 semanas de vacaciones después de 15 años de servicio es

a) cualquiera entre 2 y 5;
b) menor que 3.

5.7 Un destacado médico afirma que el 70% de las personas con cáncer de pulmón son fumadores empedernidos. Si su aseveración es correcta,

a) calcule la probabilidad de que de 10 de estos pacientes, que ingresaron recientemente a un hospital, menos de la mitad sean fumadores empedernidos;

b) calcule la probabilidad de que de 20 de estos pacientes, que ingresaron recientemente a un hospital, menos de la mitad sean fumadores empedernidos.

5.8 De acuerdo con un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, aproximadamente 60% de los consumidores de Valium en el estado de Massachusetts empezaron a consumirlo a causa de problemas psicológicos. Calcule la probabilidad de que entre los siguientes 8 consumidores entrevistados de este estado,

a) exactamente 3 comenzaron a consumir Valium por problemas psicológicos;
b) al menos 5 comenzaron a consumir Valium por problemas que no fueron psicológicos.

5.9 Al probar cierta clase de neumático para camión en un terreno accidentado, se encuentra que el 25% de los camiones no completan la prueba de recorrido sin ponchaduras. De los siguientes 15 camiones probados, calcule la probabilidad de que

a) de 3 a 6 tengan ponchaduras;
b) menos de 4 tengan ponchaduras;
c) más de 5 tengan ponchaduras.

5.10 Según un informe de la revista *Parade*, una encuesta a nivel nacional, realizada por la Universidad de Michigan con estudiantes universitarios de último año, reveló que casi 70% desaprueban el consumo diario de marihuana. Si se seleccionan 12 estudiantes de último año al azar y se les pide su opinión, calcule la probabilidad de que el número de los que desaprueban el consumo diario de marihuana sea

a) cualquiera entre 7 y 9;
b) 5 a lo sumo;
c) no menos de 8.

5.11 La probabilidad de que un paciente se recupere de una delicada operación de corazón es 0.9. ¿Cuál es la probabilidad de que exactamente 5 de los siguientes 7 pacientes intervenidos sobrevivan?

5.12 Un ingeniero de control de tráfico reporta que 75% de los vehículos que pasan por un punto de verificación son de ese estado. ¿Cuál es la probabilidad de que menos de 4 de los siguientes 9 vehículos sean de otro estado?

5.13 Un estudio a nivel nacional que examinó las actitudes hacia los antidepresivos reveló que aproximadamente 70% de los encuestados cree que "los antidepresivos en realidad no curan nada, sólo disfrazan el problema real". De acuerdo con este estudio, ¿cuál es la probabilidad de que al menos 3 de las siguientes 5 personas seleccionadas al azar tengan esta opinión?

5.14 El porcentaje de victorias que consiguió el equipo de baloncesto los Toros de Chicago para pasar a las finales en la temporada 1996-97 fue de 87.7. Redondee 87.7 a 90 para poder utilizar la tabla A.1.

a) ¿Cuál es la probabilidad de que los Toros logren una victoria aplastante (4-0) en la serie final de 7 juegos?
b) ¿Cuál es la probabilidad de que los Toros ganen la serie inicial?
c) ¿Qué suposición importante se hace al responder los incisos *a*) y *b*)?

5.15 Se sabe que 60% de los ratones inoculados con un suero quedan protegidos contra cierta enfermedad. Si se inoculan 5 ratones, calcule la probabilidad de que

a) ninguno contraiga la enfermedad;
b) menos de 2 contraigan la enfermedad;
c) más de 3 contraigan la enfermedad.

5.16 Suponga que los motores de un avión operan de forma independiente y que tienen una probabilidad de falla de 0.4. Se supone que un avión tiene un vuelo seguro si funcionan al menos la mitad de sus motores. Si un avión tiene 4 motores y otro tiene 2, ¿cuál de los dos tiene la probabilidad más alta de un vuelo exitoso?

5.17 Si X representa el número de personas del ejercicio 5.13 que creen que los antidepresivos no curan sino que sólo disfrazan el problema real, calcule la media y la varianza de X si se seleccionan al azar 5 personas.

5.18 *a*) ¿Cuántos de los 15 camiones del ejercicio 5.9 esperaría que tuvieran ponchaduras?
b) ¿Cuál es la varianza del número de ponchaduras de los 15 camiones? ¿Qué significado tiene eso?

5.19 Un estudiante que conduce hacia su escuela encuentra un semáforo, el cual permanece verde por 35 segundos, amarillo cinco segundos y rojo 60 segundos. Suponga que toda la semana el estudiante recorre el camino a la escuela entre las 8:00 y las 8:30 a.m. Sea X_1 el número de veces que encuentra una luz verde, X_2 el número de veces que encuentra una luz amarilla y X_3 el número de veces que encuentra una luz roja. Calcule la distribución conjunta de X_1, X_2 y X_3.

5.20 Según el diario *USA Today* (18 de marzo de 1997), de 4 millones de integrantes de la fuerza laboral, 5.8% resultó positivo en una prueba de drogas. De los que dieron positivo, 22.5% consumían cocaína y 54.4% consumían marihuana.

a) ¿Cuál es la probabilidad de que de 10 trabajadores que dieron positivo, 2 sean usuarios de cocaína, 5 de marihuana y 3 de otras drogas?
b) ¿Cuál es la probabilidad de que de 10 trabajadores que dieron positivo, todos sean consumidores de marihuana?
c) ¿Cuál es la probabilidad de que de 10 trabajadores que dieron positivo, ninguno consuma cocaína?

5.21 La superficie de un tablero circular para dardos tiene un pequeño círculo central llamado diana y 20 regiones en forma de rebanada de pastel numeradas del 1 al 20. Asimismo, cada una de estas regiones está dividida en tres partes, de manera que una persona que lanza un dardo que cae en un número específico obtiene una puntuación igual al valor del número, el doble del número o el triple de éste, dependiendo de en cuál de las tres partes caiga el dardo. Si una persona tiene una probabilidad de 0.01 de acertar a la diana, una probabilidad de 0.10 de acertar un doble, una probabilidad de 0.05 de acertar un triple y una probabilidad de 0.02 de no acertar al tablero, ¿cuál es la probabilidad de que 7 lanzamientos den como resultado ninguna diana, ningún triple, dos dobles y una vez fuera del tablero?

5.22 De acuerdo con la teoría genética, cierta cruza de conejillos de Indias tendrá crías rojas, negras y blancas en la proporción 8:4:4. Calcule la probabilidad de que de 8 crías, 5 sean rojas, 2 negras y 1 blanca.

5.23 Las probabilidades de que un delegado llegue a cierta convención en avión, autobús, automóvil o tren son de 0.4, 0.2, 0.3 y 0.1, respectivamente. ¿Cuál es la probabilidad de que, de 9 delegados que asisten a esta convención seleccionados al azar, 3 lleguen en avión, 3 en autobús, 1 en automóvil y 2 en tren?

5.24 Un ingeniero de seguridad afirma que sólo 40% de los trabajadores utilizan cascos de seguridad cuando comen en el lugar de trabajo. Suponga que esta afirmación es cierta y calcule la probabilidad de que 4 de 6 trabajadores elegidos al azar utilicen sus cascos mientras comen en el lugar de trabajo.

5.25 Suponga que para un embarque muy grande de circuitos integrados, la probabilidad de que falle cualquiera de ellos es de 0.10. Suponga que se cumplen los supuestos en que se basan las distribuciones binomiales y calcule la probabilidad de que en una muestra aleatoria de 20 fallen, a lo sumo, 3 chips integrados.

5.26 Suponga que 6 de 10 accidentes automovilísticos se deben principalmente a que no se respeta el límite de velocidad y calcule la probabilidad de que, de 8 accidentes automovilísticos, 6 se deban principalmente a una violación del límite de velocidad

a) mediante el uso de la fórmula para la distribución binomial;
b) usando la tabla A.1.

5.27 Si una bombilla fluorescente tiene una probabilidad de 0.9 de tener una vida útil de al menos 800 horas, calcule las probabilidades de que, de 20 bombillas fluorescentes,

a) exactamente 18 tengan una vida útil de al menos 800 horas;
b) al menos 15 tengan una vida útil de al menos 800 horas;
c) al menos 2 *no* tengan una vida útil de al menos 800 horas.

5.28 Un fabricante sabe que, en promedio, 20% de los tostadores eléctricos producidos requerirá reparaciones durante el primer año posterior a su venta. Suponga que se seleccionan al azar 20 tostadores y calcule los números x y y adecuados tales que

a) la probabilidad de que al menos x de ellos requieran reparaciones sea menor que 0.5;
b) la probabilidad de que al menos y de ellos *no* requieran reparaciones sea mayor que 0.8.

5.3 Distribución hipergeométrica

La manera más simple de ver la diferencia entre la distribución binomial de la sección 5.2 y la distribución hipergeométrica consiste en observar la forma en que se realiza el muestreo. Los tipos de aplicaciones de la distribución hipergeométrica son muy similares a los de la distribución binomial. Nos interesa el cálculo de probabilidades para el número de observaciones que caen en una categoría específica. Sin embargo, la distribución binomial requiere que los ensayos sean independientes. Por consiguiente, si se aplica esta distribución, digamos, tomando muestras de un lote de artículos (barajas, lotes de artículos producidos), el muestreo se debe efectuar **reemplazando** cada artículo después de observarlo. Por otro lado, la distribución hipergeométrica no requiere independencia y se basa en el muestreo que se realiza **sin reemplazo**.

Las aplicaciones de la distribución hipergeométrica se encuentran en muchos campos, sobre todo en el muestreo de aceptación, las pruebas electrónicas y los controles de calidad. Evidentemente, en muchos de estos campos el muestreo se realiza a expensas del artículo que se prueba; es decir, el artículo se destruye, por lo que no se puede

reemplazar en la muestra. Por consiguiente, el muestreo sin reemplazo es necesario. Utilizaremos un caso simple con barajas para nuestro primer ejemplo.

Si deseamos calcular la probabilidad de obtener 3 cartas rojas en 5 extracciones de una baraja ordinaria de 52 cartas, la distribución binomial de la sección 5.2 no se aplica a menos que cada carta se reemplace y que el paquete se revuelva antes de extraer la siguiente carta. Para resolver el problema del muestreo sin reemplazo volvamos a plantear el problema. Si se sacan 5 cartas al azar, nos interesa la probabilidad de seleccionar 3 cartas rojas de las 26 disponibles y 2 de las 26 cartas negras de que dispone la baraja. Hay $\binom{26}{3}$ formas de seleccionar 3 cartas rojas, y para cada una de estas formas podemos elegir 2 cartas negras de $\binom{26}{2}$ maneras. Por lo tanto, el número total de formas de seleccionar 3 cartas rojas y 2 negras en 5 extracciones es el producto $\binom{26}{3}\binom{26}{2}$. El número total de formas de seleccionar cualesquiera 5 cartas de las 52 disponibles es $\binom{52}{5}$. En consecuencia, la probabilidad de seleccionar 5 cartas sin reemplazo, de las cuales 3 sean rojas y 2 negras está dada por

$$\frac{\binom{26}{3}\binom{26}{2}}{\binom{52}{5}} = \frac{(26!/3!23!)(26!/2!24!)}{52!/5!47!} = 0.3251.$$

En general, nos interesa la probabilidad de seleccionar x éxitos de los k artículos considerados éxitos y $n - x$ fracasos de los $N - k$ artículos que se consideran fracasos cuando una muestra aleatoria de tamaño n se selecciona de N artículos. Esto se conoce como un **experimento hipergeométrico**; es decir, aquel que posee las siguientes dos propiedades:

1. De un lote de N artículos se selecciona una muestra aleatoria de tamaño n sin reemplazo.
2. k de los N artículos se pueden clasificar como éxitos y $N - k$ se clasifican como fracasos.

El número X de éxitos de un experimento hipergeométrico se denomina **variable aleatoria hipergeométrica**. En consecuencia, la distribución de probabilidad de la variable hipergeométrica se conoce como **distribución hipergeométrica**, y sus valores se denotan con $h(x; N, n, k)$, ya que dependen del número de éxitos k en el conjunto N del que seleccionamos n artículos.

Distribución hipergeométrica en el muestreo de aceptación

Como en el caso de la distribución binomial, la distribución hipergeométrica se aplica en el muestreo de aceptación, donde se toman muestras del material o las partes de los lotes con el fin de determinar si se acepta o no el lote completo.

Ejemplo 5.8: Una parte específica que se utiliza como dispositivo de inyección se vende en lotes de 10. El productor considera que el lote es aceptable si no tiene más de un artículo defectuoso. Un plan de muestreo incluye un muestreo aleatorio y la prueba de 3 de cada 10 partes. Si ninguna de las 3 está defectuosa, se acepta el lote. Comente acerca de la utilidad de este plan.

***Solución*:** Supongamos que el lote es verdaderamente **inaceptable** (es decir, que 2 de cada 10 partes están defectuosas). La probabilidad de que el plan de muestreo considere que el lote aceptable es

$$P(X = 0) = \frac{\binom{2}{0}\binom{8}{3}}{\binom{10}{3}} = 0.467.$$

Por consiguiente, si el lote es realmente inaceptable porque 2 partes están defectuosas, este plan de muestreo permitirá que se acepte aproximadamente 47% de las veces. Como resultado, este plan debería considerarse inadecuado. ∎

Hagamos una generalización para calcular una fórmula para $h(x; N, n, k)$. El número total de muestras de tamaño n elegidas de N artículos es $\binom{N}{n}$. Se supone que estas muestras tienen la misma probabilidad. Hay $\binom{k}{x}$ formas de seleccionar x éxitos de los k disponibles, y por cada una de estas formas podemos elegir $n - x$ fracasos en formas $\binom{N-k}{n-x}$. De esta manera, el número total de muestras favorables entre las $\binom{N}{n}$ muestras posibles, está dado por $\binom{k}{x}\binom{N-k}{n-x}$. En consecuencia, tenemos la siguiente definición.

Distribución hipergeométrica

La distribución de probabilidad de la variable aleatoria hipergeométrica X, el número de éxitos en una muestra aleatoria de tamaño n que se selecciona de N artículos, en los que k se denomina **éxito** y $N - k$ **fracaso**, es

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad \max\{0, n - (N - k)\} \leq x \leq \min\{n, k\}.$$

El rango de x puede determinarse mediante los tres coeficientes binomiales en la definición, donde x y $n - x$ no son más que k y $N - k$; respectivamente; y ambos no pueden ser menores que 0. Por lo general, cuando tanto k (el número de éxitos) como $N - k$ (el número de fracasos) son mayores que el tamaño de la muestra n, el rango de una variable aleatoria hipergeométrica será $x = 0, 1, ..., n$.

Ejemplo 5.9: Lotes con 40 componentes cada uno que contengan 3 o más defectuosos se consideran inaceptables. El procedimiento para obtener muestras del lote consiste en seleccionar 5 componentes al azar y rechazar el lote si se encuentra un componente defectuoso. ¿Cuál es la probabilidad de, que en la muestra, se encuentre exactamente un componente defectuoso, si en todo el lote hay 3 defectuosos?

***Solución*:** Si utilizamos la distribución hipergeométrica con $n = 5$, $N = 40$, $k = 3$ y $x = 1$, encontramos que la probabilidad de obtener un componente defectuoso es

$$h(1; 40, 5, 3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011.$$

De nueva cuenta este plan no es adecuado porque sólo 30% de las veces detecta un lote malo (con 3 componentes defectuosos). ∎

Teorema 5.2: La media y la varianza de la distribución hipergeométrica $h(x; N, n, k)$ son

$$\mu = \frac{nk}{N} \text{ y } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right).$$

La demostración para la media se muestra en el apéndice A.24.

Ejemplo 5.10: Volvamos a investigar el ejemplo 3.4 de la página 83. La finalidad de este ejemplo fue ilustrar el concepto de una variable aleatoria y el espacio muestral correspondiente. En el ejemplo tenemos un lote de 100 artículos, de los cuales 12 están defectuosos. ¿Cuál es la probabilidad de que haya 3 defectuosos en una muestra de 10?

***Solución*:** Si utilizamos la función de probabilidad hipergeométrica, tenemos

$$h(3; 100, 10, 12) = \frac{\binom{12}{3}\binom{88}{7}}{\binom{100}{10}} = 0.08.$$

Ejemplo 5.11: Calcule la media y la varianza de la variable aleatoria del ejemplo 5.9, y después utilice el teorema de Chebyshev para interpretar el intervalo $\mu \pm 2\sigma$.

***Solución*:** Como el ejemplo 5.9 fue un experimento hipergeométrico con $N = 40$, $n = 5$ y $k = 3$, usando el teorema 5.2, tenemos

$$\mu = \frac{(5)(3)}{40} = \frac{3}{8} = 0.375,$$

y

$$\sigma^2 = \left(\frac{40-5}{39}\right)(5)\left(\frac{3}{40}\right)\left(1-\frac{3}{40}\right) = 0.3113.$$

Si calculamos la raíz cuadrada de 0.3113, encontramos que $\sigma = 0.558$. Por lo tanto, el intervalo que se requiere es $0.375 \pm (2)(0.558)$, o de -0.741 a 1.491. El teorema de Chebyshev establece que el número de componentes defectuosos que se obtienen cuando, de un lote de 40 componentes, se seleccionan 5 al azar, de los cuales 3 están defectuosos, tiene una probabilidad de al menos 3/4 de caer entre -0.741 y 1.491. Esto es, al menos tres cuartas partes de las veces los 5 componentes incluirán menos de 2 defectuosos.

Relación con la distribución binomial

En este capítulo examinamos varias distribuciones discretas importantes que tienen diversas aplicaciones. Muchas de estas distribuciones se relacionan bien entre sí. El estudiante novato debería tener una clara comprensión de tales relaciones. Existe una relación interesante entre las distribuciones hipergeométrica y binomial. Como se esperaría, si n es pequeña comparada con N, la naturaleza de los N artículos cambia muy poco en cada prueba. Así, cuando n es pequeña en comparación con N, se puede utilizar una distribución binomial para aproximar la distribución hipergeométrica. De hecho, por regla general la aproximación es buena cuando $n/N \leq 0.05$.

Por lo tanto, la cantidad k/N desempeña el papel del parámetro binomial p y, como consecuencia, la distribución binomial se podría considerar una versión de población grande de la distribución hipergeométrica. La media y la varianza entonces se obtienen de las fórmulas

$$\mu = np = \frac{nk}{N} \text{ y } \sigma^2 = npq = n \cdot \frac{k}{N}\left(1-\frac{k}{N}\right).$$

Al comparar estas fórmulas con las del teorema 5.2, vemos que la media es la misma, mientras que la varianza difiere por un factor de corrección de $(N-n)/(N-1)$, que es insignificante cuando n es pequeña en relación con N.

Ejemplo 5.12: Un fabricante de neumáticos para automóvil reporta que de un cargamento de 5000 piezas que se mandan a un distribuidor local, 1000 están ligeramente manchadas. Si se compran al azar 10 de estos neumáticos al distribuidor, ¿cuál es la probabilidad de que exactamente 3 estén manchados?

***Solución*:** Como $N = 5000$ es grande con respecto a la muestra de tamaño $n = 10$, nos aproximaremos a la probabilidad deseada usando la distribución binomial. La probabilidad de obtener un neumático manchado es 0.2. Por lo tanto, la probabilidad de obtener exactamente 3 manchados es

$$h(3; 5000, 10, 1000) \approx b(3; 10, 0.2) = 0.8791 - 0.6778 = 0.2013.$$

Por otro lado, la probabilidad exacta es $h(3; 5000, 10, 1000) = 0.2015$. ■

La distribución hipergeométrica se puede extender para tratar el caso donde los N artículos se pueden dividir en k celdas $A_1, A_2, ..., A_k$ con a_1 elementos en la primera celda, a_2 en la segunda,..., a_k elementos en la k-ésima celda. Lo que nos interesa ahora es la probabilidad de que una muestra aleatoria de tamaño n produzca x_1 elementos de A_1, x_2 elementos de A_2, ... , y x_k elementos de A_k. Representemos esta probabilidad mediante

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n).$$

Para obtener una fórmula general observamos que el número total de muestras de tamaño n que se pueden elegir a partir de N artículos es aún $\binom{N}{n}$. Hay $\binom{a_1}{x_1}$ formas de seleccionar x_1 artículos de los que hay en A_1, y para cada uno de éstos podemos elegir x_2 de los de A_2 en $\binom{a_2}{x_2}$ formas. Por lo tanto, podemos seleccionar x_1 artículos de A_1 y x_2 artículos de A_2 en $\binom{a_1}{x_1}\binom{a_2}{x_2}$ formas. Si continuamos de esta forma, podemos seleccionar todos los n artículos que constan de x_1 de A_1, x_2 de A_2,..., y x_k de A_k en

$$\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k} \quad \text{formas.}$$

La distribución de probabilidad que se requiere se define ahora como sigue.

Distribución hipergeométrica multivariada

Si N artículos se pueden dividir en las k celdas $A_1, A_2, ..., A_k$ con $a_1, a_2, ..., a_k$ elementos, respectivamente, entonces la distribución de probabilidad de las variables aleatorias X_1, X_2,..., X_k, que representan el número de elementos que se seleccionan de $A_1, A_2, ..., A_k$ en una muestra aleatoria de tamaño n, es

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}},$$

con $\sum_{i=1}^{k} x_i = n$ y $\sum_{i=1}^{k} a_i = N$.

Ejemplo 5.13: Se usa un grupo de 10 individuos para un estudio de caso biológico. El grupo contiene 3 personas con sangre tipo O, 4 con sangre tipo A y 3 con tipo B. ¿Cuál es la probabilidad de que una muestra aleatoria de 5 contenga 1 persona con sangre tipo O, 2 personas con tipo A y 2 personas con tipo B?

***Solución*:** Si se utiliza la extensión de la distribución hipergeométrica con $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $a_1 = 3$, $a_2 = 4$, $a_3 = 3$, $N = 10$ y $n = 5$, vemos que la probabilidad que se desea es

$$f(1, 2, 2; 3, 4, 3, 10, 5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14}.$$

■

Ejercicios

5.29 El dueño de una casa planta 6 bulbos seleccionados al azar de una caja que contiene 5 bulbos de tulipán y 4 de narciso. ¿Cuál es la probabilidad de que plante 2 bulbos de narciso y 4 de tulipán?

5.30 Para evitar la detección en la aduana, un viajero coloca 6 comprimidos con narcóticos en una botella que contiene 9 píldoras de vitamina que aparentemente son similares. Si el oficial de la aduana selecciona 3 de las tabletas al azar para su análisis, ¿cuál es la probabilidad de que el viajero sea arrestado por posesión ilegal de narcóticos?

5.31 Se selecciona al azar un comité de 3 personas a partir de 4 médicos y 2 enfermeras. Escriba una fórmula para la distribución de probabilidad de la variable aleatoria X que representa el número de médicos en el comité. Calcule $P(2 \leq X \leq 3)$.

5.32 De un lote de 10 misiles, se seleccionan 4 al azar y se disparan. Si el lote contiene 3 misiles defectuosos que no pueden dispararse, ¿cuál es la probabilidad de que
a) los 4 puedan dispararse?
b) a lo sumo fallen 2?

5.33 Si de una baraja ordinaria de 52 cartas, se toman 7 y se reparten, ¿cuál es la probabilidad de que
a) exactamente 2 de ellas sean cartas de figuras?
b) al menos 1 de ellas sea una reina?

5.34 ¿Cuál es la probabilidad de que una camarera se rehúse a servir bebidas alcohólicas a sólo 2 menores si verifica al azar 5 identificaciones de 9 estudiantes, de los cuales 4 son menores de edad?

5.35 Una empresa está interesada en evaluar su procedimiento de inspección actual para embarques de 50 artículos idénticos. El procedimiento consiste en tomar una muestra de 5 artículos y aceptar el embarque si no se encuentran más de 2 defectuosos. ¿Qué proporción de embarques con 20% de artículos defectuosos se aceptará?

5.36 Una empresa de manufactura utiliza un esquema de aceptación para los artículos de una línea de producción antes de que se embarquen. El plan tiene dos etapas. Se preparan cajas de 25 artículos para su embarque y se prueba una muestra de 3 en busca de defectuosos. Si se encuentra alguno defectuoso, se regresa toda la caja para verificar el 100% de ellos. Si no se encuentran artículos defectuosos, la caja se embarca.
a) ¿Cuál es la probabilidad de que se embarque una caja que contiene 3 defectuosos?
b) ¿Cuál es la probabilidad de que se regrese para su revisión una caja que contenga sólo un artículo defectuoso?

5.37 Suponga que la empresa fabricante del ejercicio 5.36 decide cambiar su esquema de aceptación. Con el nuevo esquema un inspector toma un artículo al azar, lo inspecciona y después lo regresa a la caja; un segundo inspector hace lo mismo. Finalmente, un tercer inspector lleva a cabo el mismo procedimiento. Si cualquiera de los tres encuentra un artículo defectuoso, la caja no se embarca. Responda los incisos del ejercicio 5.36 con este nuevo plan.

5.38 De los 150 empleados de hacienda en una ciudad grande, sólo 30 son mujeres. Suponga que se eligen al azar 10 de los empleados para que proporcionen asesoría gratuita sobre declaraciones de impuestos a los residentes de esta ciudad; utilice la aproximación binomial a la distribución hipergeométrica para calcular la probabilidad de que se seleccionen al menos 3 mujeres.

5.39 Una ciudad vecina considera entablar una demanda de anexión en contra de una subdivisión del condado de 1200 residencias. Si los ocupantes de la mitad de las residencias objetan la anexión, ¿cuál es la probabilidad de que en una muestra aleatoria de 10 residencias al menos 3 estén a favor de la anexión?

5.40 Se estima que 4000 de los 10,000 residentes con derecho al voto de una ciudad están en contra de un nuevo impuesto sobre las ventas. Si se seleccionan al azar 15 votantes y se les pide su opinión, ¿cuál es la probabilidad de que a lo sumo 7 estén a favor del nuevo impuesto?

5.41 Una encuesta a nivel nacional, realizada por la Universidad de Michigan a 17,000 estudiantes universitarios de último año, revela que casi 70% desaprueba el consumo diario de marihuana. Si se seleccionan al azar 18 de tales estudiantes y se les pide su opinión, ¿cuál es la probabilidad de que más de 9 pero menos de 14 desaprueben el consumo de marihuana?

5.42 Calcule la probabilidad de que si le toca una mano de bridge de 13 cartas, ésta incluya 5 espadas, 2 corazones, 3 diamantes y 3 tréboles.

5.43 Un club de estudiantes extranjeros tiene como miembros a 2 canadienses, 3 japoneses, 5 italianos y 2 alemanes. Si se selecciona al azar un comité de 4, calcule la probabilidad de que
a) todas las nacionalidades estén representadas;
b) todas las nacionalidades estén representadas, excepto la italiana.

5.44 Una urna contiene 3 bolas verdes, 2 azules y 4 rojas. Calcule la probabilidad de que, en una muestra aleatoria de 5 bolas, se seleccionen las 2 bolas azules y al menos una roja.

5.45 A menudo los biólogos que estudian un ambiente específico etiquetan y liberan a sujetos con el fin de estimar el tamaño de la población o la prevalencia de ciertas características en ella. Los biólogos capturan a 10 animales de una especie que se piensa extinta (o casi extinta), los etiquetan y los liberan en cierta región. Después de un periodo seleccionan en la región una muestra aleatoria de 15 animales de ese tipo. ¿Cuál es la probabilidad de que 5 de los animales seleccionados estén etiquetados, si hay 25 animales de este tipo en la región?

5.46 Una empresa grande tiene un sistema de inspección para los lotes de compresores pequeños que compra a los vendedores. Un lote típico contiene 15 compresores. En el sistema de inspección se selecciona una muestra aleatoria de 5 compresores para someterlos a prueba. Suponga que en el lote de 15 hay 2 defectuosos.

a) ¿Cuál es la probabilidad de que en una muestra determinada haya un compresor defectuoso?

b) ¿Cuál es la probabilidad de que la inspección descubra los 2 compresores defectuosos?

5.47 Una fuerza de tareas gubernamental sospecha que algunas fábricas infringen los reglamentos federales contra la contaminación ambiental en lo que se refiere a la descarga de cierto tipo de producto. Veinte empresas están bajo sospecha pero no todas se pueden inspeccionar. Suponga que 3 de las empresas infringen los reglamentos.

a) ¿Cuál es la probabilidad de que si se inspeccionan 5 empresas no se encuentre ninguna infracción?

b) ¿Cuál es la probabilidad de que la inspección de 5 empresas descubra a 2 que infringen el reglamento?

5.48 Una máquina llena 10,000 latas de bebida gaseosa por hora, de entre las cuales 300 resultan con el líquido incompleto. Cada hora se elige al azar una muestra de 30 latas y se verifica el número de onzas de gaseosa que contiene cada una. Denote con X el número de latas seleccionadas con llenado insuficiente. Encuentre la probabilidad de encontrar al menos una de las latas muestreadas con llenado insuficiente.

5.4 Distribuciones binomial negativa y geométrica

Consideremos un experimento con las mismas propiedades de un experimento binomial, sólo que en este caso las pruebas se repetirán hasta que ocurra un número *fijo* de éxitos. Por lo tanto, en vez de encontrar la probabilidad de x éxitos en n pruebas, donde n es fija, ahora nos interesa la probabilidad de que ocurra el k-ésimo éxito en la x-ésima prueba. Los experimentos de este tipo se llaman **experimentos binomiales negativos**.

Como ejemplo, considere el uso de un medicamento que se sabe que es eficaz en el 60% de los casos en que se utiliza. El uso del medicamento se considerará un éxito si proporciona algún grado de alivio al paciente. Nos interesa calcular la probabilidad de que el quinto paciente que experimente alivio sea el séptimo paciente en recibir el medicamento en una semana determinada. Si designamos un éxito con E y un fracaso con F, un orden posible para alcanzar el resultado que se desea es $EFEEEFE$, que ocurre con la siguiente probabilidad

$$(0.6)\,(0.4)\,(0.6)\,(0.6)\,(0.6)\,(0.4)\,(0.6) = (0.6)^5\,(0.4)^2.$$

Podríamos listar todos los posibles ordenamientos reacomodando las F y las E, con excepción del último resultado, que debe ser el quinto éxito. El número total de ordenamientos posibles es igual al número de particiones de los primeros 6 ensayos en 2 grupos con dos fracasos asignados a un grupo y 4 éxitos asignados al otro grupo. Esto se puede realizar en $\binom{6}{4} = 15$ formas mutuamente excluyentes. Por lo tanto, si X representa el resultado en el que ocurre el quinto éxito, entonces

$$P(X = 7) = \binom{6}{4}(0.6)^5(0.4)^2 = 0.1866.$$

¿Cuál es la variable aleatoria binomial negativa?

El número X de ensayos necesarios para generar k éxitos en un experimento binomial negativo se denomina **variable aleatoria binomial negativa** y su distribución de probabi-

lidad se llama **distribución binomial negativa**. Dado que sus probabilidades dependen del número de éxitos deseados y de la probabilidad de un éxito en un ensayo dado, denotaremos ambas probabilidades con el símbolo $b^*(x; k, p)$. Para obtener la fórmula general para $b^*(x; k, p)$, considere la probabilidad de un éxito en el x-ésimo ensayo precedido por $k-1$ éxitos y $x-k$ fracasos en un orden específico. Como los ensayos son independientes podemos multiplicar todas las probabilidades que corresponden a cada resultado deseado. La probabilidad de que ocurra un éxito es p y la probabilidad de que ocurra un fracaso es $q = 1 - p$. Por lo tanto, la probabilidad para el orden específico, que termina en un éxito, es

$$p^{k-1}q^{x-k}p = p^k q^{x-k}.$$

El número total de puntos muestrales en el experimento que termina en un éxito, después de la ocurrencia de $k-1$ éxitos y $x-k$ fracasos en cualquier orden, es igual al número de particiones de $x-1$ ensayos en dos grupos con $k-1$ éxitos, que corresponden a un grupo, y $x-k$ fracasos, que corresponden al otro grupo. Este número se especifica con el término $\binom{x-1}{k-1}$, cada uno es mutuamente excluyente y tiene las mismas probabilidades de ocurrir $p^k q^{x-k}$. Obtenemos la fórmula general multiplicando $p^k q^{x-k}$ por $\binom{x-1}{k-1}$.

Distribución binomial negativa

Si ensayos independientes repetidos pueden dar como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número del ensayo en el que ocurre el k-ésimo éxito, es

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \ldots$$

Ejemplo 5.14: En la serie de campeonato de la NBA (National Basketball Association), el equipo que gane 4 de 7 juegos será el ganador. Suponga que los equipos A y B se enfrentan en los juegos de campeonato y que el equipo A tiene una probabilidad de 0.55 de ganarle al equipo B.

a) ¿Cuál es la probabilidad de que el equipo A gane la serie en 6 juegos?

b) ¿Cuál es la probabilidad de que el equipo A gane la serie?

c) Si ambos equipos se enfrentaran en la eliminatoria de una serie regional y el triunfador fuera el que ganara 3 de 5 juegos, ¿cuál es la probabilidad de que el equipo A gane la serie?

Solución: *a*) $b^*(6; 4, 0.55) = \binom{5}{3}0.55^4(1-0.55)^{6-4} = 0.1853.$

b) P(el equipo A gana la serie de campeonato) es

$$\begin{aligned} &b^*(4; 4, 0.55) + b^*(5; 4, 0.55) + b^*(6; 4, 0.55) + b^*(7; 4, 0.55) \\ &\quad = 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083. \end{aligned}$$

c) P(el equipo A gana la eliminatoria) es

$$\begin{aligned} &b^*(3; 3, 0.55) + b^*(4; 3, 0.55) + b^*(5; 3, 0.55) \\ &\quad = 0.1664 + 0.2246 + 0.2021 = 0.5931. \end{aligned}$$

La distribución binomial negativa deriva su nombre del hecho de que cada término de la expansión de $p^k(1-q)^{-k}$ corresponde a los valores de $b^*(x; k, p)$ para $x = k$, $k+1$, $k+2$,.... Si consideramos el caso especial de la distribución binomial negativa, donde $k = 1$, tenemos una distribución de probabilidad para el número de ensayos que se requieren para un solo éxito. Un ejemplo sería lanzar una moneda hasta que salga una cara. Nos podemos interesar en la probabilidad de que la primera cara resulte en el cuarto lanzamiento. En este caso la distribución binomial negativa se reduce a la forma

$$b^*(x; 1, p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots$$

Como los términos sucesivos constituyen una progresión geométrica, se acostumbra referirse a este caso especial como **distribución geométrica** y denotar sus valores con $g(x; p)$.

Distribución geométrica

Si pruebas independientes repetidas pueden tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número de la prueba en el que ocurre el primer éxito, es

$$g(x; p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots$$

Ejemplo 5.15: Se sabe que en cierto proceso de fabricación uno de cada 100 artículos, en promedio, resulta defectuoso. ¿Cuál es la probabilidad de que el quinto artículo que se inspecciona, en un grupo de 100, sea el primer defectuoso que se encuentra?

***Solución*:** Si utilizamos la distribución geométrica con $x = 5$ y $p = 0.01$, tenemos

$$g(5; 0.01) = (0.01)(0.99)^4 = 0.0096.$$

Ejemplo 5.16: En "momentos ajetreados" un conmutador telefónico está muy cerca de su límite de capacidad, por lo que los usuarios tienen dificultad para hacer sus llamadas. Sería interesante saber cuántos intentos serían necesarios para conseguir un enlace telefónico. Suponga que la probabilidad de conseguir un enlace durante un momento ajetreado es $p = 0.05$. Nos interesa conocer la probabilidad de que se necesiten 5 intentos para enlazar con éxito una llamada.

***Solución*:** Si utilizamos la distribución geométrica con $x = 5$ y $p = 0.05$, obtenemos

$$P(X = x) = g(5; 0.05) = (0.05)(0.95)^4 = 0.041.$$

Muy a menudo, en aplicaciones que tienen que ver con la distribución geométrica, la media y la varianza son importantes. Se puede ver esto en el ejemplo 5.16, en donde el número *esperado* de llamadas necesario para lograr un enlace es muy importante. A continuación se establecen, sin demostración, la media y la varianza de la distribución geométrica.

Teorema 5.3: La media y la varianza de una variable aleatoria que sigue la distribución geométrica son

$$\mu = \frac{1}{p} \text{ y } \sigma^2 = \frac{1-p}{p^2}.$$

Aplicaciones de las distribuciones binomial negativa y geométrica

Las áreas de aplicación de las distribuciones binomial negativa y geométrica serán evidentes cuando nos enfoquemos en los ejemplos de esta sección y en los ejercicios que se dedican a tales distribuciones al final de la sección 5.5. En el caso de la distribución geométrica, el ejemplo 5.16 describe una situación en que los ingenieros o administradores intentan determinar cuán ineficiente es un sistema de conmutación telefónica durante periodos ajetreados. En este caso es evidente que los ensayos que ocurren antes de un éxito representan un costo. Si hay una alta probabilidad de que se requieran varios intentos antes de lograr conectarse, entonces se debería rediseñar el sistema.

Las aplicaciones de la distribución binomial negativa son similares por naturaleza. Supongamos que los intentos son costosos en algún sentido y que *ocurren en secuencia*. La alta probabilidad de que se requiera un número "grande" de intentos para experimentar un número fijo de éxitos no es benéfica ni para el científico ni para el ingeniero. Considere los escenarios de los ejercicios de repaso 5.90 y 5.91. En el ejercicio 5.91 el perforador define cierto nivel de éxitos perforando diferentes sitios en secuencia para encontrar petróleo. Si sólo se han hecho 6 intentos en el momento en que se experimenta el segundo éxito, parecería que las utilidades superan de forma considerable la inversión en que se incurre para la perforación.

5.5 Distribución de Poisson y proceso de Poisson

Los experimentos que producen valores numéricos de una variable aleatoria X, el número de resultados que ocurren durante un intervalo de tiempo determinado o en una región específica, se denominan **experimentos de Poisson**. El intervalo de tiempo puede ser de cualquier duración, como un minuto, un día, una semana, un mes o incluso un año. Por ejemplo, un experimento de Poisson podría generar observaciones para la variable aleatoria X que representa el número de llamadas telefónicas por hora que recibe una oficina, el número de días que una escuela permanece cerrada debido a la nieve durante el invierno o el número de juegos suspendidos debido a la lluvia durante la temporada de béisbol. La región específica podría ser un segmento de recta, una área, un volumen o quizá una pieza de material. En tales casos X podría representar el número de ratas de campo por acre, el número de bacterias en un cultivo dado o el número de errores mecanográficos por página. Un experimento de Poisson se deriva del **proceso de Poisson** y tiene las siguientes propiedades:

Propiedades del proceso de Poisson

1. El número de resultados que ocurren en un intervalo o región específica es independiente del número que ocurre en cualquier otro intervalo de tiempo o región del espacio disjunto. De esta forma vemos que el proceso de Poisson no tiene memoria.
2. La probabilidad de que ocurra un solo resultado durante un intervalo de tiempo muy corto o en una región pequeña es proporcional a la longitud del intervalo o al tamaño de la región, y no depende del número de resultados que ocurren fuera de este intervalo de tiempo o región.
3. La probabilidad de que ocurra más de un resultado en tal intervalo de tiempo corto o que caiga en tal región pequeña es insignificante.

El número X de resultados que ocurren durante un experimento de Poisson se llama **variable aleatoria de Poisson** y su distribución de probabilidad se llama **distribu-**

ción de Poisson. El número medio de resultados se calcula a partir de $\mu = \lambda t$, donde t es el "tiempo", la "distancia", el "área" o el "volumen" específicos de interés. Como las probabilidades dependen de λ, denotaremos la tasa de ocurrencia de los resultados con $p(x; \lambda t)$. La derivación de la fórmula para $p(x; \lambda t)$, que se basa en las tres propiedades de un proceso de Poisson que se listaron antes, está fuera del alcance de este texto. La siguiente fórmula se utiliza para calcular probabilidades de Poisson.

Distribución de Poisson

La distribución de probabilidad de la variable aleatoria de Poisson X, la cual representa el número de resultados que ocurren en un intervalo de tiempo dado o región específicos y se denota con t, es

$$p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

donde λ es el número promedio de resultados por unidad de tiempo, distancia, área o volumen y $e = 2.71828\ldots$

La tabla A.2 contiene las sumatorias de la probabilidad de Poisson

$$P(r; \lambda t) = \sum_{x=0}^{r} p(x; \lambda t),$$

para valores selectos de λt que van de 0.1 a 18.0 Ilustramos el uso de esta tabla con los siguientes dos ejemplos.

Ejemplo 5.17: Durante un experimento de laboratorio el número promedio de partículas radiactivas que pasan a través de un contador en un milisegundo es 4. ¿Cuál es la probabilidad de que entren 6 partículas al contador en un milisegundo dado?

***Solución*:** Al usar la distribución de Poisson con $x = 6$ y $\lambda t = 4$, y al remitirnos a la tabla A.2, tenemos que

$$p(6; 4) = \frac{e^{-4}4^6}{6!} = \sum_{x=0}^{6} p(x; 4) - \sum_{x=0}^{5} p(x; 4) = 0.8893 - 0.7851 = 0.1042.$$

Ejemplo 5.18: El número promedio de camiones-tanque que llega cada día a cierta ciudad portuaria es 10. Las instalaciones en el puerto pueden alojar a lo sumo 15 camiones-tanque por día. ¿Cuál es la probabilidad de que en un día determinado lleguen más de 15 camiones y se tenga que rechazar algunos?

***Solución*:** Sea X el número de camiones-tanque que llegan cada día. Entonces, usando la tabla A.2, tenemos

$$P(X > 15) = 1 - P(X \leq 15) = 1 - \sum_{x=0}^{15} p(x; 10) = 1 - 0.9513 = 0.0487.$$

Como la distribución binomial, la distribución de Poisson se utiliza para control de calidad, aseguramiento de calidad y muestreo de aceptación. Además, ciertas distribuciones continuas importantes que se usan en la teoría de confiabilidad y en la teoría de colas dependen del proceso de Poisson. Algunas de estas distribuciones se analizan y desarrollan en el capítulo 6. El siguiente teorema acerca de la variable aleatoria de Poisson se presenta en el apéndice A.25.

Teorema 5.4: Tanto la media como la varianza de la distribución de Poisson $p(x; \lambda t)$ son λt.

Naturaleza de la función de probabilidad de Poisson

Al igual que muchas distribuciones discretas y continuas, la forma de la distribución de Poisson se vuelve cada vez más simétrica, incluso con forma de campana, a medida que la media se hace más grande. Una ilustración de esto son las gráficas de la función de probabilidad para $\mu = 0.1$, $\mu = 2$ y finalmente $\mu = 5$ que se muestran en la figura 5.1. Observe cómo se acercan a la simetría cuando μ se vuelve tan grande como 5. Con la distribución binomial ocurre algo parecido, como se ilustrará más adelante en este texto.

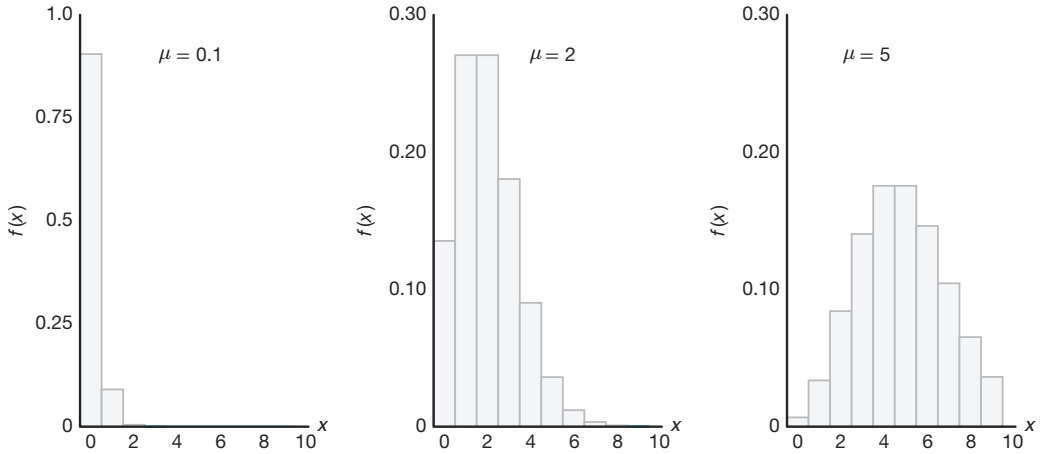


Figura 5.1: Funciones de densidad de Poisson para diferentes medias.

Aproximación de una distribución binomial por medio de una distribución de Poisson

A partir de los tres principios del proceso de Poisson debería ser evidente que la distribución de Poisson se relaciona con la distribución binomial. Aunque la de Poisson por lo general se aplica en problemas de espacio y tiempo, como se ilustra con los ejemplos 5.17 y 5.18, se podría considerar como una forma limitante de la distribución binomial. En el caso de la distribución binomial, si n es bastante grande y p es pequeña, las condiciones comienzan a simular las implicaciones *de espacio* o *tiempo continuos* del proceso de Poisson. La independencia entre las pruebas de Bernoulli en el caso binomial es consistente con la segunda propiedad del proceso de Poisson. Permitir que el parámetro p se acerque a cero se relaciona con la tercera propiedad del proceso de Poisson. De hecho, si n es grande y p es cercana a 0, se puede usar la distribución de Poisson, con $\mu = np$, para aproximar probabilidades binomiales. Si p es cercana a 1, aún podemos utilizar la distribución de Poisson para aproximar probabilidades binomiales intercambiando lo que definimos como éxito y fracaso, por lo tanto, cambiando p a un valor cercano a 0.

Teorema 5.5: Sea X una variable aleatoria binomial con distribución de probabilidad $b(x; n, p)$. Cuando $n \to \infty$, $p \to 0$, y $np \overset{n\to\infty}{\longrightarrow} \mu$ permanece constante,

$$b(x; n, p) \overset{n\to\infty}{\longrightarrow} p(x; \mu).$$

Ejemplo 5.19: En cierta fábrica los accidentes ocurren con muy poca frecuencia. Se sabe que la probabilidad de un accidente en cualquier día dado es de 0.005, y que los accidentes son independientes entre sí.

a) ¿Cuál es la probabilidad de que en un día de cualquier periodo determinado de 400 días ocurra un accidente?

b) ¿Cuál es la probabilidad de que ocurra un accidente a lo sumo en tres días de tal periodo?

***Solución*:** Sea X una variable aleatoria binomial con $n = 400$ y $p = 0.005$. Por consiguiente, $np = 2$. Si utilizamos la aproximación de Poisson,

a) $P(X = 1) = e^{-2}2^1 = 0.271$ y

b) $P(X \leq 3) = \sum_{x=0}^{3} e^{-2}2^x/x! = 0.857.$

Ejemplo 5.20: En un proceso de fabricación donde se manufacturan productos de vidrio ocurren defectos o burbujas, lo cual ocasionalmente hace que la pieza ya no se pueda vender. Se sabe que, en promedio, 1 de cada 1000 artículos producidos tiene una o más burbujas. ¿Cuál es la probabilidad de que una muestra aleatoria de 8000 tenga menos de 7 artículos con burbujas?

***Solución*:** Se trata básicamente de un experimento binomial con $n = 8000$ y $p = 0.001$. Como p es muy cercana a cero y n es bastante grande, haremos la aproximación con la distribución de Poisson utilizando

$$\mu = (8000)(0.001) = 8.$$

Por lo tanto, si X representa el número de burbujas, tenemos

$$P(X < 7) = \sum_{x=0}^{6} b(x; 8000, 0.001) \approx p(x; 8) = 0.3134.$$

Ejercicios

5.49 La probabilidad de que una persona que vive en cierta ciudad tenga un perro es de 0.3. Calcule la probabilidad de que la décima persona entrevistada al azar en esa ciudad sea la quinta que tiene un perro.

5.50 Calcule la probabilidad de que una persona que lanza una moneda obtenga

a) la tercera cara en el séptimo lanzamiento;

b) la primera cara en el cuarto lanzamiento.

5.51 Tres personas lanzan una moneda legal y el disparejo paga los cafés. Si todas las monedas tienen el mismo resultado, se lanzan de nuevo. Calcule la probabilidad de que se necesiten menos de 4 lanzamientos.

5.52 Un científico inocula a varios ratones, uno a la vez, el virus que produce una enfermedad, hasta que encuentra a 2 que contraen la enfermedad. Si la probabilidad de contraer la enfermedad es de 1/6, ¿cuál es la probabilidad de que tenga que inocular a 8 ratones?

5.53 Un estudio de un inventario determina que, en promedio, el número de veces al día que se solicita un artículo específico en un almacén es 5. ¿Cuál es la probabilidad de que en un día determinado este artículo se pida

a) más de 5 veces?

b) ninguna vez?

5.54 De acuerdo con un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, Estados Unidos, casi dos terceras partes de los 20 millones de personas que consumen Valium son mujeres. Suponga que esta cifra es una estimación válida y calcule la probabilidad de que en un determinado día la quinta prescripción de Valium que da un médico sea

a) la primera prescripción de Valium para una mujer;

b) la tercera prescripción de Valium para una mujer.

5.55 La probabilidad de que una persona que estudia la carrera de piloto privado apruebe el examen escrito para obtener la licencia es de 0.7. Calcule la probabilidad de que cierto estudiante apruebe el examen
a) en el tercer intento;
b) antes del cuarto intento.

5.56 En cierto crucero ocurren, en promedio, 3 accidentes de tránsito al mes. ¿Cuál es la probabilidad de que en cualquier determinado mes en este crucero
a) ocurran exactamente 5 accidentes?
b) ocurran menos de 3 accidentes?
c) ocurran al menos 2 accidentes?

5.57 Un escritor de libros comete, en promedio, dos errores de procesamiento de texto por página en el primer borrador de su libro. ¿Cuál es la probabilidad de que en la siguiente página cometa
a) 4 o más errores?
b) ningún error?

5.58 Cierta área del este de Estados Unidos resulta afectada, en promedio, por 6 huracanes al año. Calcule la probabilidad de que para cierto año esta área resulte afectada por
a) menos de 4 huracanes;
b) cualquier cantidad entre 6 y 8 huracanes.

5.59 Suponga que la probabilidad de que una determinada persona crea un rumor acerca de las transgresiones de cierta actriz famosa es de 0.8. ¿Cuál es la probabilidad de que
a) la sexta persona que escuche este rumor sea la cuarta en creerlo?
b) la tercera persona que escuche este rumor sea la primera en creerlo?

5.60 Se estima que el número promedio de ratas de campo por acre, en un campo de 5 acres de trigo, es 12. Calcule la probabilidad de que se encuentren menos de 7 ratas de campo
a) en un acre dado;
b) en 2 de los siguientes 3 acres que se inspeccionen.

5.61 Suponga que, en promedio, una persona en 1000 comete un error numérico al preparar su declaración de impuestos. Si se seleccionan 10,000 formas al azar y se examinan, calcule la probabilidad de que 6, 7 u 8 de las formas contengan un error.

5.62 Se sabe que la probabilidad de que un estudiante de preparatoria no pase la prueba de escoliosis (curvatura de la espina dorsal) es de 0.004. De los siguientes 1875 estudiantes que se revisan en búsqueda de escoliosis, calcule la probabilidad de que
a) menos de 5 no pasen la prueba;
b) 8, 9 o 10 no pasen la prueba.

5.63 Calcule la media y la varianza de la variable aleatoria X del ejercicio 5.58, que representa el número de huracanes que afectan cada año a cierta área del este de Estados Unidos.

5.64 Calcule la media y la varianza de la variable aleatoria X del ejercicio 5.61, que representa el número de personas, de cada 10,000, que comete un error al preparar su declaración de impuestos.

5.65 Un fabricante de automóviles se preocupa por una falla en el mecanismo de freno de un modelo específico. En raras ocasiones la falla puede causar una catástrofe al manejarlo a alta velocidad. La distribución del número de automóviles por año que experimentará la catástrofe es una variable aleatoria de Poisson con $\lambda = 5$.
a) ¿Cuál es la probabilidad de que, a lo sumo, 3 automóviles por año de ese modelo específico sufran una catástrofe?
b) ¿Cuál es la probabilidad de que más de un automóvil por año experimente una catástrofe?

5.66 Los cambios en los procedimientos de los aeropuertos requieren una planeación considerable. Los índices de llegadas de los aviones son factores importantes que deben tomarse en cuenta. Suponga que los aviones pequeños llegan a cierto aeropuerto, de acuerdo con un proceso de Poisson, con una frecuencia de 6 por hora. De esta manera, el parámetro de Poisson para las llegadas en un periodo de horas es $\mu = 6t$.
a) ¿Cuál es la probabilidad de que lleguen exactamente 4 aviones pequeños durante un periodo de una hora?
b) ¿Cuál es la probabilidad de que lleguen al menos 4 durante un periodo de una hora?
c) Si definimos un día laboral como de 12 horas, ¿cuál es la probabilidad de que al menos 75 aviones pequeños lleguen durante un día laboral?

5.67 Se supone que el número de clientes que llegan por hora a ciertas instalaciones de servicio automotriz sigue una distribución de Poisson con media $\lambda = 7$.
a) Calcule la probabilidad de que lleguen más de 10 clientes en un periodo de dos horas.
b) ¿Cuál es el número medio de llegadas durante un periodo de 2 horas?

5.68 Considere el ejercicio 5.62. ¿Cuál es el número promedio de estudiantes que no pasan la prueba?

5.69 La probabilidad de que una persona muera al contraer una infección viral es de 0.001. De los siguientes 4000 infectados con el virus, ¿cuál es el número promedio que morirá?

5.70 Una empresa compra lotes grandes de cierta clase de dispositivo electrónico. Utiliza un método que rechaza el lote completo si en una muestra aleatoria de 100 unidades se encuentran 2 o más unidades defectuosas.

a) ¿Cuál es el número promedio de unidades defectuosas que se encuentran en una muestra de 100 unidades si el lote tiene 1% de unidades defectuosas?

b) ¿Cuál es la varianza?

5.71 Se sabe que para cierto tipo de alambre de cobre ocurren, en promedio, 1.5 fallas por milímetro. Si se supone que el número de fallas es una variable aleatoria de Poisson, ¿cuál es la probabilidad de que no ocurran fallas en cierta parte de un alambre que tiene 5 milímetros de longitud? ¿Cuál es el número promedio de fallas en alguna parte de un alambre que tiene 5 milímetros de longitud?

5.72 Los baches en ciertas carreteras pueden ser un problema grave y requieren reparación constantemente. Con un tipo específico de terreno y mezcla de concreto la experiencia sugiere que hay, en promedio, 2 baches por milla después de cierta cantidad de uso. Se supone que el proceso de Poisson se aplica a la variable aleatoria "número de baches".

a) ¿Cuál es la probabilidad de que no aparezca más de un bache en un tramo de una milla?

b) ¿Cuál es la probabilidad de que no aparezcan más de 4 baches en un tramo determinado de 5 millas?

5.73 En ciudades grandes los administradores de los hospitales se preocupan por el flujo de personas en las salas de urgencias. En un hospital específico de una ciudad grande el personal disponible no puede alojar el flujo de pacientes cuando hay más de 10 casos de emergencia en una hora determinada. Se supone que la llegada de los pacientes sigue un proceso de Poisson y los datos históricos sugieren que, en promedio, llegan 5 emergencias cada hora.

a) ¿Cuál es la probabilidad de que en una hora determinada el personal no pueda alojar el flujo de pacientes?

b) ¿Cuál es la probabilidad de que, durante un turno de 3 horas, lleguen más de 20 emergencias?

5.74 Se sabe que 3% de las personas a las que se les revisa el equipaje en un aeropuerto lleva objetos cuestionables. ¿Cuál es la probabilidad de que una serie de 15 personas cruce sin problemas antes de que se atrape a una con un objeto cuestionable? ¿Cuál es el número esperado de personas que pasarán antes de que se detenga a una?

5.75 La tecnología cibernética ha generado un ambiente donde los "robots" funcionan con el uso de microprocesadores. La probabilidad de que un robot falle durante cualquier turno de 6 horas es de 0.10. ¿Cuál es la probabilidad de que un robot funcione a lo sumo 5 turnos antes de fallar?

5.76 Se sabe que la tasa de rechazo en las encuestas telefónicas es de aproximadamente 20%. Un reportaje del periódico indica que 50 personas respondieron a una encuesta antes de que una se rehusara a participar.

a) Comente acerca de la validez del reportaje. Utilice una probabilidad en su argumento.

b) ¿Cuál es el número esperado de personas encuestadas antes de que una se rehúse a responder?

Ejercicios de repaso

5.77 Durante un proceso de producción, cada día se seleccionan al azar 15 unidades de la línea de ensamble para verificar el porcentaje de artículos defectuosos. A partir de información histórica se sabe que la probabilidad de tener una unidad defectuosa es de 0.05. Cada vez que se encuentran dos o más unidades defectuosas en la muestra de 15, el proceso se detiene. Este procedimiento se utiliza para proporcionar una señal en caso de que aumente la probabilidad de unidades defectuosas.

a) ¿Cuál es la probabilidad de que en un día determinado se detenga el proceso de producción? (Suponga 5% de unidades defectuosas).

b) Suponga que la probabilidad de una unidad defectuosa aumenta a 0.07. ¿Cuál es la probabilidad de que en cualquier día no se detenga el proceso de producción?

5.78 Se considera utilizar una máquina automática de soldadura para un proceso de producción. Antes de comprarla se probará para verificar si tiene éxito en 99% de sus soldaduras. Si no es así, se considerará que no es eficiente. La prueba se llevará a cabo con un prototipo que requiere hacer 100 soldaduras. La máquina se aceptará para la producción sólo si no falla en más de 3 soldaduras.

a) ¿Cuál es la probabilidad de que se rechace una buena máquina?

b) ¿Cuál es la probabilidad de que se acepte una máquina ineficiente que solde bien el 95% de las veces?

5.79 Una agencia de renta de automóviles en un aeropuerto local tiene 5 Ford, 7 Chevrolet, 4 Dodge, 3 Honda y 4 Toyota disponibles. Si la agencia selecciona al azar 9 de estos automóviles para transportar delega-

dos desde el aeropuerto hasta el centro de convenciones de la ciudad, calcule la probabilidad de que rente 2 Ford, 3 Chevrolet, 1 Dodge, 1 Honda y 2 Toyota.

5.80 En un centro de mantenimiento que recibe llamadas de servicio de acuerdo con un proceso de Poisson entran, en promedio, 2.7 llamadas por minuto. Calcule la probabilidad de que

a) no entren más de 4 llamadas en cualquier minuto;
b) entren menos de 2 llamadas en cualquier minuto;
c) entren más de 10 llamadas en un periodo de 5 minutos.

5.81 Una empresa de electrónica afirma que la proporción de unidades defectuosas de cierto proceso es de 5%. Un comprador sigue el procedimiento estándar de inspeccionar 15 unidades elegidas al azar de un lote grande. En una ocasión específica el comprador encuentra 5 unidades defectuosas.

a) ¿Cuál es la probabilidad de que esto ocurra, si es correcta la afirmación de que el 5% de los productos son defectuosos?
b) ¿Cómo reaccionaría usted si fuera el comprador?

5.82 Un dispositivo electrónico de conmutación falla ocasionalmente, pero se considera que es satisfactorio si, en promedio, no comete más de 0.20 errores por hora. Se elige un periodo particular de 5 horas para probarlo. Si durante este periodo no ocurre más de un error, se considera que el funcionamiento del dispositivo es satisfactorio.

a) ¿Cuál es la probabilidad de que, con base en la prueba, se considere que un dispositivo no funciona satisfactoriamente cuando en realidad sí lo hace? Suponga que se trata de un proceso de Poisson.
b) ¿Cuál es la probabilidad de que un dispositivo se considere satisfactorio cuando, de hecho, el número medio de errores que comete es 0.25? De nuevo suponga que se trata de un proceso de Poisson.

5.83 Una empresa por lo general compra lotes grandes de cierta clase de dispositivo electrónico. Utiliza un método que rechaza el lote completo si encuentra 2 o más unidades defectuosas en una muestra aleatoria de 100 unidades.

a) ¿Cuál es la probabilidad de que el método rechace un lote que tiene un 1% de unidades defectuosas?
b) ¿Cuál es la probabilidad de que acepte un lote que tiene 5% de unidades defectuosas?

5.84 El propietario de una farmacia local sabe que, en promedio, llegan a su farmacia 100 personas por hora.

a) Calcule la probabilidad de que en un periodo determinado de 3 minutos nadie entre a la farmacia.
b) Calcule la probabilidad de que en un periodo dado de 3 minutos entren más de 5 personas a la farmacia.

5.85 a) Suponga que lanza 4 dados. Calcule la probabilidad de obtener al menos un 1.
b) Suponga que lanza 2 dados 24 veces. Calcule la probabilidad de obtener al menos uno (1, 1), es decir, un "ojos de serpiente".

5.86 Suponga que de 500 billetes de lotería que se venden, 200 le dan a ganar al comprador al menos el costo del billete. Ahora suponga que usted compra 5 billetes. Calcule la probabilidad de ganar al menos el costo de 3 billetes.

5.87 Las imperfecciones en los tableros de circuitos y los microcircuitos de computadora se prestan para un análisis estadístico. Un tipo particular de tablero contiene 200 diodos y la probabilidad de que falle alguno es de 0.03.

a) ¿Cuál es el número promedio de fallas en los diodos?
b) ¿Cuál es la varianza?
c) El tablero funciona si no tiene diodos defectuosos. ¿Cuál es la probabilidad de que un tablero funcione?

5.88 El comprador potencial de un motor particular requiere (entre otras cosas) que éste encienda 10 veces consecutivas. Suponga que la probabilidad de que encienda es de 0.990. Suponga que los resultados de intentos de encendido son independientes.

a) ¿Cuál es la probabilidad de que el posible comprador acepte el motor después de sólo 10 encendidos?
b) ¿Cuál es la probabilidad de que se tenga que intentar encenderlo 12 veces durante el proceso de aceptación?

5.89 El esquema de aceptación para comprar lotes que contienen un número grande de baterías consiste en probar no más de 75 baterías seleccionadas al azar y rechazar el lote completo si falla una sola batería. Suponga que la probabilidad de encontrar una que falle es de 0.001.

a) ¿Cuál es la probabilidad de que se acepte un lote?
b) ¿Cuál es la probabilidad de que se rechace un lote en la vigésima prueba?
c) ¿Cuál es la probabilidad de que se rechace en 10 o menos pruebas?

5.90 Una empresa que perfora pozos petroleros opera en varios sitios y su éxito o fracaso es independiente de un sitio a otro. Suponga que la probabilidad de éxito en cualquier sitio específico es de 0.25.

a) ¿Cuál es la probabilidad de que un perforador barrene 10 sitios y tenga un éxito?
b) El perforador se declarará en bancarrota si tiene que perforar 10 veces antes de que ocurra el primer éxito. ¿Cuáles son las perspectivas de bancarrota del perforador?

5.91 Considere la información del ejercicio de repaso 5.90. El perforador cree que "dará en el clavo" si logra el segundo éxito durante o antes del sexto intento. ¿Cuál es la probabilidad de que el perforador "dé en el clavo"?

5.92 Una pareja decide que continuará procreando hijos hasta tener dos hombres. Suponiendo que $P(\text{hombre}) = 0.5$, ¿cuál es la probabilidad de que su segundo niño sea su cuarto hijo?

5.93 Por los investigadores se sabe que una de cada 100 personas es portadora de un gen que lleva a la herencia de cierta enfermedad crónica. En una muestra aleatoria de 1000 individuos, ¿cuál es la probabilidad de que menos de 7 individuos porten el gen? Utilice la aproximación de Poisson. Nuevamente con la aproximación de Poisson, determine cuál es el número promedio aproximado de personas, de cada 1000, que portan el gen.

5.94 Un proceso de fabricación produce piezas para componentes electrónicos. Se supone que la probabilidad de que una pieza salga defectuosa es de 0.01. Durante una prueba de esta suposición se obtiene una muestra al azar de 500 artículos y se encuentran 15 defectuosos.

a) ¿Cuál es su respuesta ante la suposición de que 1% de las piezas producidas salen defectuosas? Asegúrese de acompañar su comentario con un cálculo de probabilidad.
b) Suponiendo que 1% de las piezas producidas salen con defecto, ¿cuál es la probabilidad de que sólo se encuentren 3 defectuosas?
c) Resuelva de nueva cuenta los incisos *a*) y *b*) utilizando la aproximación de Poisson.

5.95 Un proceso de manufactura produce artículos en lotes de 50. Se dispone de planes de muestreo en los cuales los lotes se apartan periódicamente y se someten a cierto tipo de inspección. Por lo general se supone que la proporción de artículos defectuosos que resultan del proceso es muy pequeña. Para la empresa también es importante que los lotes que contengan artículos defectuosos sean un evento raro. El plan actual de inspección consiste en elegir lotes al azar, obtener muestras periódicas de 10 en 50 artículos de un lote y, si ninguno de los muestreados está defectuoso, no se realizan acciones.

a) Suponga que se elige un lote al azar y 2 de cada 50 artículos tienen defecto. ¿Cuál es la probabilidad de que al menos uno en la muestra de 10 del lote esté defectuoso?
b) A partir de su respuesta en el inciso *a*), comente sobre la calidad de este plan de muestreo.
c) ¿Cuál es el número promedio de artículos defectuosos encontrados por cada 10 artículos de la muestra?

5.96 Considere la situación del ejercicio de repaso 5.95. Se ha determinado que el plan de muestreo debería ser lo suficientemente amplio como para que haya una probabilidad alta, digamos de 0.9, de que si hay tantos como 2 artículos defectuosos en el lote de 50 que se muestrea, al menos uno se encuentre en el muestreo. Con tales restricciones, ¿cuántos de los 50 artículos deberían muestrearse?

5.97 La seguridad nacional requiere que la tecnología de defensa sea capaz de detectar proyectiles o misiles ofensivos. Para que este sistema de defensa sea exitoso, se requieren múltiples pantallas de radar. Suponga que se usarán tres pantallas independientes y que la probabilidad de que cualquiera detecte un misil ofensivo es de 0.8. Es evidente que si ninguna pantalla detecta un misil ofensivo, el sistema no funciona y requiere mejorarse.

a) ¿Cuál es la probabilidad de que ninguna de las pantallas detecte un misil ofensivo?
b) ¿Cuál es la probabilidad de que sólo una de las pantallas detecte el misil?
c) ¿Cuál es la probabilidad de que al menos 2 de las 3 pantallas detecten el misil?

5.98 Suponga que es importante que el sistema general de defensa contra misiles sea lo más perfecto posible.

a) Suponga que la calidad de las pantallas es la que se indica en el ejercicio de repaso 5.97. ¿Cuántas se requieren, entonces, para asegurar que la probabilidad de que el misil pase sin ser detectado sea de 0.0001?
b) Suponga que se decide utilizar sólo 3 pantallas e intentar mejorar la capacidad de detección de las mismas. ¿Cuál debe ser la eficacia individual de las pantallas (es decir, la probabilidad de detección), para alcanzar la eficacia que se requiere en el inciso *a*?

5.99 Regrese al ejercicio de repaso 5.95*a*. Vuelva a calcular la probabilidad usando la distribución binomial. Comente su respuesta.

5.100 En cierto departamento universitario de estadística hay dos vacantes. Cinco personas las solicitan; dos de ellas tienen experiencia con modelos lineales y una tiene experiencia con probabilidad aplicada. Al comité de selección se le indicó elegir a los 2 aspirantes aleatoriamente.

a) ¿Cuál es la probabilidad de que los 2 elegidos sean los que tienen experiencia con modelos lineales?
b) ¿Cuál es la probabilidad de que, de los 2 elegidos, uno tenga experiencia con modelos lineales y el otro con probabilidad aplicada?

5.101 El fabricante de un triciclo para niños ha recibido quejas porque su producto tiene defecto en los frenos. De acuerdo con el diseño del producto y muchas pruebas preliminares, se determinó que la probabilidad del tipo de defecto reportado era 1 en 10,000 (es decir, de 0.0001). Después de una minuciosa investigación de las quejas se determinó que durante cierto periodo se eligieron aleatoriamente 200 artículos de la producción, de los cuales 5 tuvieron frenos defectuosos.

a) Comente sobre la afirmación de "uno en 10,000" del fabricante. Utilice un argumento probabilístico. Use la distribución binomial para sus cálculos.

b) Repita el inciso *a* utilizando la aproximación de Poisson.

5.102 Proyecto de grupo: Separe la clase en dos grupos aproximadamente del mismo tamaño. Cada uno de los estudiantes del grupo 1 lanzará una moneda 10 veces (n_1) y contará el número de caras resultantes. Cada uno de los estudiantes del grupo 2 lanzará una moneda 40 veces (n_2) y también contará el número de caras obtenidas. Los miembros de cada grupo deben calcular de manera individual la proporción de caras observadas, que es una estimación de p, la probabilidad de obtener una cara. De esta manera, habrá un conjunto de valores de p_1 (del grupo 1) y un conjunto de valores de p_2 (del grupo 2). Todos los valores de p_1 y p_2 son estimaciones de 0.5, que es el valor verdadero de la probabilidad de obtener una cara de una moneda legal.

a) ¿Cuál conjunto de valores se acerca con mayor consistencia a 0.5, el de p_1 o el de p_2? Considere la demostración del teorema 5.1 de la página 147 con respecto a las estimaciones del parámetro $p = 0.5$. Los valores de p_1 se obtuvieron con $n = n_1 = 10$ y los valores de p_2 se obtuvieron con $n = n_2 = 40$. Si se utiliza la notación de la demostración, las estimaciones están dadas por

$$p_1 = \frac{x_1}{n_1} = \frac{I_1 + \cdots + I_{n_1}}{n_1},$$

donde $I_1,\ldots, I_{n_1}$ son ceros y unos y $n_1 = 10$, y

$$p_2 = \frac{x_2}{n_2} = \frac{I_1 + \cdots + I_{n_2}}{n_2},$$

donde $I_1,\ldots,I_{n_2}$ son nuevamente ceros y unos y $n_2 = 40$.

b) Remítase nuevamente al teorema 5.1 y demuestre que

$$E(p_1) = E(p_2) = p = 0.5.$$

c) Demuestre que $\sigma^2_{p_1} = \frac{\sigma^2_{X_1}}{n_1}$ es 4 veces el valor de $\sigma^2_{p_2} = \frac{\sigma^2_{X_2}}{n_2}$. Explique, además, por qué los valores de p_2 del grupo 2 se acercan con mayor consistencia al valor verdadero, $p = 0.5$, que los valores de p_1 del grupo 1.

Aprenderá mucho más sobre la estimación de parámetros a partir del capítulo 9. Ahí pondremos más énfasis en la importancia de la media y la varianza de un estimador de un parámetro.

5.6 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Las distribuciones discretas estudiadas en este capítulo ocurren con mucha frecuencia en los escenarios de la ingeniería, así como en los de las ciencias biológicas y físicas. Es evidente que los ejemplos y los ejercicios sugieren esto. Los planes de muestreo industrial y muchas de las decisiones en ingeniería se basan en las distribuciones binomial y de Poisson, así como en la distribución hipergeométrica. Mientras que las distribuciones binomial negativa y geométrica se utilizan en menor grado, también tienen aplicaciones. En específico, una variable aleatoria binomial negativa se puede ver como una mezcla de variables aleatorias gamma y de Poisson (la distribución gamma se estudiará en el capítulo 6).

A pesar de las múltiples aplicaciones que estas distribuciones tienen en la vida real, podrían utilizarse de manera incorrecta, a menos que el científico sea prudente y cuidadoso. Desde luego, cualquier cálculo de probabilidad para las distribuciones que se estudiaron en este capítulo se realiza bajo el supuesto de que se conoce el valor del parámetro. Las aplicaciones en el mundo real a menudo resultan en un valor del parámetro que se puede "desplazar" debido a factores que son difíciles de controlar en el proceso,

o debido a intervenciones en el proceso que no se han tomado en cuenta. Por ejemplo, en el ejercicio de repaso 5.77 se utilizó "información histórica"; sin embargo, ¿el proceso actual es el mismo que aquel en que se recabaron los datos históricos? El uso de la distribución de Poisson tiene incluso más posibilidades de enfrentar esta dificultad. Por ejemplo, en el ejercicio de repaso 5.80 las preguntas de los incisos *a*, *b* y *c* se basan en el uso de $\mu = 2.7$ llamadas por minuto. Con base en los registros históricos éste es el número de llamadas que se reciben "en promedio". Pero en ésta y muchas otras aplicaciones de la distribución de Poisson hay momentos desocupados y momentos ajetreados, de manera que se espera que haya momentos en que las condiciones para el proceso de Poisson parezcan cumplirse, cuando en realidad no lo hacen. Por consiguiente, los cálculos de probabilidad podrían ser incorrectos. En el caso de la distribución binomial, la condición que podría fallar en ciertas aplicaciones (además de la falta de constancia de *p*) es la suposición de independencia, estipulando que los experimentos de Bernoulli son independientes.

Una de las aplicaciones incorrectas más célebres de la distribución binomial ocurrió en la temporada de béisbol de 1961, cuando Mickey Mantle y Roger Maris se enfrascaron en una batalla amistosa por romper el récord de todos los tiempos de 60 jonrones establecido por Babe Ruth. Un famoso artículo de una revista predijo, con base en la teoría de la probabilidad, que Mantle rompería el récord. La predicción estaba fundamentada en un cálculo de probabilidad en el que se utilizó la distribución binomial. El error clásico cometido fue la estimación del parámetro *p* (uno para cada jugador) con base en la frecuencia histórica relativa de jonrones a lo largo de la carrera de los 2 jugadores. Maris, a diferencia de Mantle, no había sido un jonronero prodigioso antes de 1961, de manera que su estimado de *p* fue bastante bajo. Como resultado de esto se determinó que Mantle tenía más probabilidades que Maris de romper el récord, pero quien logró romperlo al final fue este último.

Capítulo 6

Algunas distribuciones continuas de probabilidad

6.1 Distribución uniforme continua

Una de las distribuciones continuas más simples de la estadística es la **distribución uniforme continua**. Esta distribución se caracteriza por una función de densidad que es "plana", por lo cual la probabilidad es uniforme en un intervalo cerrado, digamos $[A, B]$. Aunque las aplicaciones de la distribución uniforme continua no son tan abundantes como las de otras distribuciones que se presentan en este capítulo, es apropiado para el principiante que comience esta introducción a las distribuciones continuas con la distribución uniforme.

Distribución uniforme

La función de densidad de la variable aleatoria uniforme continua X en el intervalo $[A, B]$ es

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B, \\ 0, & \text{en otro caso.} \end{cases}$$

La función de densidad forma un rectángulo con base $B - A$ y **altura constante** $\frac{1}{B-A}$. Como resultado, la distribución uniforme a menudo se conoce como **distribución rectangular**. Sin embargo, observe que el intervalo no siempre es cerrado: $[A, B]$; también puede ser (A, B). En la figura 6.1 se muestra la función de densidad para una variable aleatoria uniforme en el intervalo $[1, 3]$.

Resulta sencillo calcular las probabilidades para la distribución uniforme debido a la naturaleza simple de la función de densidad. Sin embargo, observe que la aplicación de esta distribución se basa en el supuesto de que la probabilidad de caer en un intervalo de longitud fija dentro de $[A, B]$ es constante.

Ejemplo 6.1: Suponga que el tiempo máximo que se puede reservar una sala de conferencias grande de cierta empresa son cuatro horas. Con mucha frecuencia tienen conferencias extensas y breves. De hecho, se puede suponer que la duración X de una conferencia tiene una distribución uniforme en el intervalo $[0, 4]$.

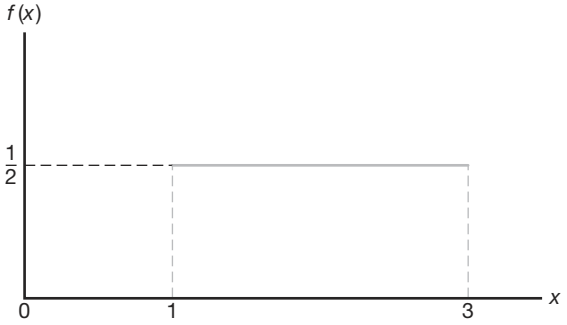


Figura 6.1: Función de densidad para una variable aleatoria en el intervalo [1, 3].

a) ¿Cuál es la función de densidad de probabilidad?

b) ¿Cuál es la probabilidad de que cualquier conferencia determinada dure al menos 3 horas?

***Solución*:** *a*) La función de densidad apropiada para la variable aleatoria X distribuida uniformemente en esta situación es

$$f(x) = \begin{cases} \frac{1}{4}, & 0 \leq x \leq 4, \\ 0, & \text{en otro caso.} \end{cases}$$

b) $P[X \geq 3] = \int_3^4 \frac{1}{4}\, dx = \frac{1}{4}$. ∎

Teorema 6.1: La media y la varianza de la distribución uniforme son

$$\mu = \frac{A + B}{2} \text{ y } \sigma^2 = \frac{(B - A)^2}{12}.$$

Las demostraciones de los teoremas se dejan al lector. Véase el ejercicio 6.1 de la página 185.

6.2 Distribución normal

La distribución de probabilidad continua más importante en todo el campo de la estadística es la **distribución normal**. Su gráfica, denominada **curva normal**, es la curva con forma de campana de la figura 6.2, la cual describe de manera aproximada muchos fenómenos que ocurren en la naturaleza, la industria y la investigación. Por ejemplo, las mediciones físicas en áreas como los experimentos meteorológicos, estudios de la precipitación pluvial y mediciones de partes fabricadas a menudo se explican más que adecuadamente con una distribución normal. Además, los errores en las mediciones científicas se aproximan muy bien mediante una distribución normal. En 1733, Abraham DeMoivre desarrolló la ecuación matemática de la curva normal, la cual sentó las bases sobre las que descansa gran parte de la teoría de la estadística inductiva. La distribución normal a menudo se denomina **distribución gaussiana** en honor de Karl Friedrich Gauss (1777-1855), quien también derivó su ecuación a partir de un estudio de errores en mediciones repetidas de la misma cantidad.

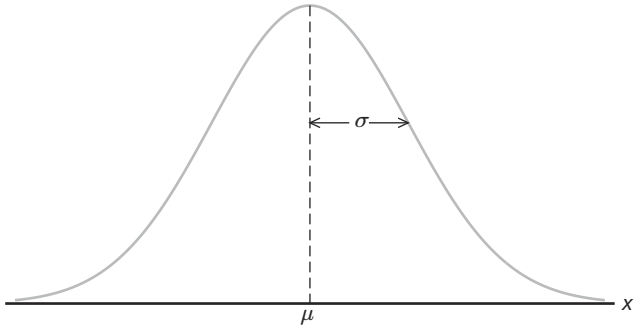


Figura 6.2: La curva normal.

Una variable aleatoria continua X que tiene la distribución en forma de campana de la figura 6.2 se denomina **variable aleatoria normal**. La ecuación matemática para la distribución de probabilidad de la variable normal depende de los dos parámetros μ y σ, su media y su desviación estándar, respectivamente. Por ello, denotamos los valores de la densidad de X por $n(x; \mu, \sigma)$.

Distribución normal

La densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty,$$

donde $\pi = 3.14159\ldots$ y $e = 2.71828\ldots$

Una vez que se especifican μ y σ, la curva normal queda determinada por completo. Por ejemplo, si $\mu = 50$ y $\sigma = 5$, entonces se pueden calcular las ordenadas $n(x; 50, 5)$ para diferentes valores de x y dibujar la curva. En la figura 6.3 aparecen dos curvas normales que tienen la misma desviación estándar pero diferentes medias. Las dos curvas son idénticas en forma, pero están centradas en diferentes posiciones a lo largo del eje horizontal.

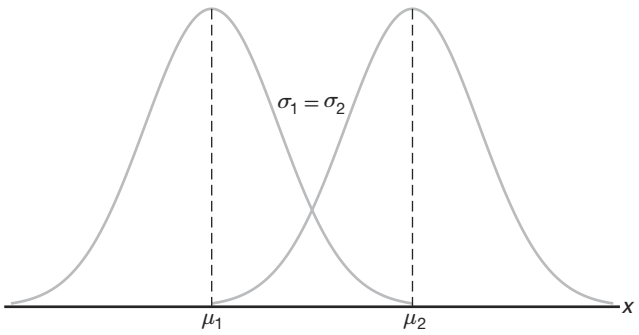


Figura 6.3: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 = \sigma_2$.

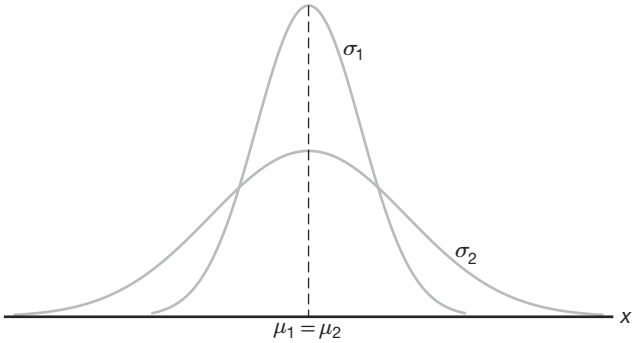


Figura 6.4: Curvas normales con $\mu_1 = \mu_2$ y $\sigma_1 < \sigma_2$.

En la figura 6.4 se muestran dos curvas normales con la misma media pero con desviaciones estándar diferentes. Aquí se observa que las dos curvas están centradas exactamente en la misma posición sobre el eje horizontal; sin embargo, la curva con la mayor desviación estándar es más baja y más extendida. Recuerde que el área bajo una curva de probabilidad debe ser igual a 1 y, por lo tanto, cuanto más variable sea el conjunto de observaciones, más baja y más ancha será la curva correspondiente.

La figura 6.5 muestra dos curvas normales que tienen diferentes medias y diferentes desviaciones estándar. Evidentemente, están centradas en posiciones diferentes sobre el eje horizontal y sus formas reflejan los dos valores diferentes de σ.

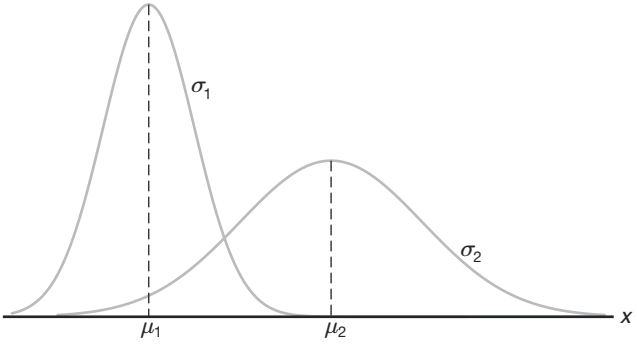


Figura 6.5: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 < \sigma_2$.

Con base en lo que observamos en las figuras 6.2 a 6.5, y en el examen de la primera y la segunda derivadas de $n(x; \mu, \sigma)$, listamos las siguientes propiedades de la curva normal:

1. La moda, que es el punto sobre el eje horizontal donde la curva tiene su punto máximo, ocurre en $x = \mu$.
2. La curva es simétrica alrededor de un eje vertical a través de la media μ.
3. La curva tiene sus puntos de inflexión en $x = \mu \pm \sigma$, es cóncava hacia abajo si $\mu - \sigma < X < \mu + \sigma$, y es cóncava hacia arriba en otro caso.

4. La curva normal se aproxima al eje horizontal de manera asintótica, conforme nos alejamos de la media en cualquier dirección.
5. El área total bajo la curva y sobre el eje horizontal es igual a uno.

Teorema 6.2: La media y la varianza de $n(x; \mu, \sigma)$ son μ y σ^2, respectivamente. Por lo tanto, la desviación estándar es σ.

***Prueba*:** Para evaluar la media primero calculamos

$$E(X - \mu) = \int_{-\infty}^{\infty} \frac{x - \mu}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Al establecer que $z = (x - \mu)/\sigma$ y $dx = \sigma\, dz$, obtenemos

$$E(X - \mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2}\, dz = 0,$$

dado que la integral anterior es una función impar de z. Al aplicar el teorema 4.5 de la página 128 concluimos que

$$E(X) = \mu$$

La varianza de la distribución normal es dada por

$$E[(X - \mu)^2] = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\frac{1}{2}[(x-\mu)/\sigma]^2} dx.$$

De nuevo, al establecer que $z = (x - \mu)/\sigma$ y $dx = \sigma\, dz$, obtenemos

$$E[(X - \mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\, dz.$$

Al integrar por partes con $u = z$ y $dv = ze^{-z^2/2}dz$ de modo que $du = dz$ y $v = -e^{-z/2}$, encontramos que

$$E[(X - \mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}} \left(-ze^{-z^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2}\, dz \right) = \sigma^2(0 + 1) = \sigma^2.$$

Muchas variables aleatorias tienen distribuciones de probabilidad que se pueden describir de forma adecuada mediante la curva normal, una vez que se especifiquen μ y σ^2. En este capítulo supondremos que se conocen estos dos parámetros, quizás a partir de investigaciones anteriores. Más adelante haremos inferencias estadísticas cuando se desconozcan μ y σ^2 y se estimen a partir de los datos experimentales disponibles.

Anteriormente señalamos el papel que desempeña la distribución normal como una aproximación razonable de variables científicas en experimentos de la vida real. Hay otras aplicaciones de la distribución normal que el lector apreciará a medida que avance en el estudio de este libro. La distribución normal tiene muchas aplicaciones como *distribución limitante*. En ciertas condiciones, la distribución normal ofrece una buena aproximación continua a las distribuciones binomial e hipergeométrica. El caso de la aproximación a la distribución binomial se examina en la sección 6.5. En el capítulo 8 el lector aprenderá acerca de las **distribuciones muestrales**. Resulta que la distribución limitante de promedios muestrales es normal, lo que brinda una base amplia para la

inferencia estadística, que es muy valiosa para el analista de datos interesado en la estimación y prueba de hipótesis. Las teorías de áreas importantes como el análisis de varianza (capítulos 13, 14 y 15) y el control de calidad (capítulo 17) se basan en suposiciones que utilizan la distribución normal.

En la sección 6.3 se ofrecen ejemplos para demostrar cómo se utilizan las tablas de la distribución normal. En la sección 6.4 continúan los ejemplos de aplicaciones de la distribución normal.

6.3 Áreas bajo la curva normal

La curva de cualquier distribución continua de probabilidad o función de densidad se construye de manera que el área bajo la curva limitada por las dos ordenadas $x = x_1$ y $x = x_2$ sea igual a la probabilidad de que la variable aleatoria X tome un valor entre $x = x_1$ y $x = x_2$. Por consiguiente, para la curva normal de la figura 6.6,

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)\, dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx,$$

es representada por el área de la región sombreada.

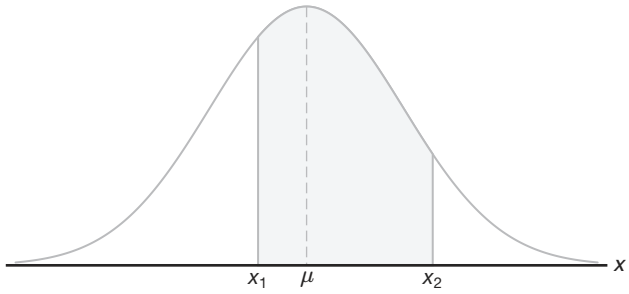


Figura 6.6: $P(x_1 < X < x_2)$ = área de la región sombreada.

En las figuras 6.3, 6.4 y 6.5 vimos cómo la curva normal depende de la media y de la desviación estándar de la distribución que se está estudiando. El área bajo la curva entre cualesquiera dos ordenadas también debe depender de los valores μ y σ. Esto es evidente en la figura 6.7, donde sombreamos las regiones que corresponden a $P(x_1 < X < x_2)$ para dos curvas con medias y varianzas diferentes. $P(x_1 < X < x_2)$, donde X es la variable aleatoria que describe la distribución A, se indica por el área sombreada más oscura debajo de la curva de A. Si X es la variable aleatoria que describe la distribución B, entonces $P(x_1 < X < x_2)$ es dada por toda la región sombreada. Evidentemente, las dos regiones sombreadas tienen tamaños diferentes; por lo tanto, la probabilidad asociada con cada distribución será diferente para los dos valores dados de X.

Existen muchos tipos de programas estadísticos que sirven para calcular el área bajo la curva normal. La dificultad que se enfrenta al resolver las integrales de funciones de densidad normal exige tabular las áreas de la curva normal para una referencia rápida. Sin embargo, sería inútil tratar de establecer tablas separadas para cada posible valor de μ y σ. Por fortuna, podemos transformar todas las observaciones de cualquier variable

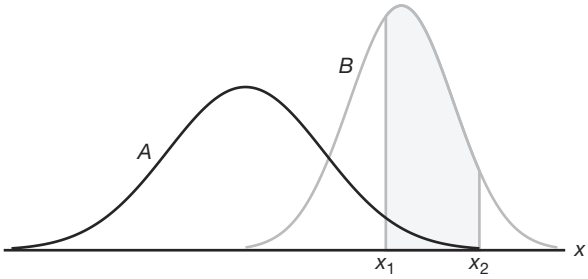


Figura 6.7: $P(x_1 < X < x_2)$ para diferentes curvas normales.

aleatoria normal X en un nuevo conjunto de observaciones de una variable aleatoria normal Z con media 0 y varianza 1. Esto se puede realizar mediante la transformación

$$Z = \frac{X - \mu}{\sigma}.$$

Siempre que X tome un valor x, el valor correspondiente de Z es dado por $z = (x - \mu)/\sigma$. Por lo tanto, si X cae entre los valores $x = x_1$ y $x = x_2$, la variable aleatoria Z caerá entre los valores correspondientes $z_1 = (x_1 - \mu)/\sigma$ y $z_2 = (x_2 - \mu)/\sigma$. En consecuencia, podemos escribir

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\,dz$$

$$= \int_{z_1}^{z_2} n(z; 0, 1)\,dz = P(z_1 < Z < z_2),$$

donde Z se considera una variable aleatoria normal con media 0 y varianza 1.

Definición 6.1: La distribución de una variable aleatoria normal con media 0 y varianza 1 se llama **distribución normal estándar**.

Las distribuciones original y transformada se ilustran en la figura 6.8. Como todos los valores de X que caen entre x_1 y x_2 tienen valores z correspondientes entre z_1 y z_2, el área bajo la curva X entre las ordenadas $x = x_1$ y $x = x_2$ de la figura 6.8 es igual al área bajo la curva Z entre las ordenadas transformadas $z = z_1$ y $z = z_2$.

Ahora hemos reducido el número requerido de tablas de áreas de curva normal a una, la de la distribución normal estándar. La tabla A.3 indica el área bajo la curva normal estándar que corresponde a $P(Z < z)$ para valores de z que van de –3.49 a 3.49. Para ilustrar el uso de esta tabla calculemos la probabilidad de que Z sea menor que 1.74. Primero, localizamos un valor de z igual a 1.7 en la columna izquierda, después nos movemos a lo largo del renglón hasta la columna bajo 0.04, donde leemos 0.9591. Por lo tanto, $P(Z < 1.74) = 0.9591$. Para calcular un valor z que corresponda a una probabilidad dada se invierte el proceso. Por ejemplo, se observa que el valor z que deja un área de 0.2148 bajo la curva a la izquierda de z es –0.79.

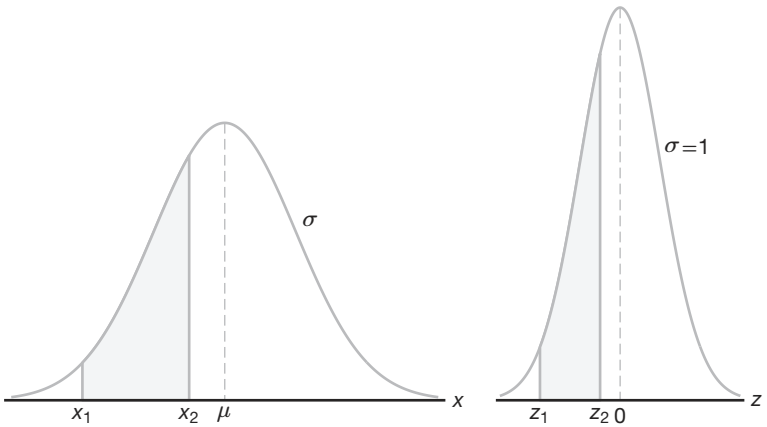


Figura 6.8: Distribuciones normales original y transformada.

Ejemplo 6.2: Dada una distribución normal estándar, calcule el área bajo la curva que se localiza

a) a la derecha de $z = 1.84$, y

b) entre $z = -1.97$ y $z = 0.86$.

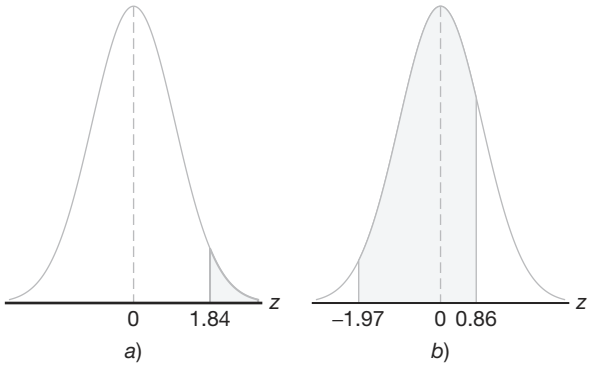


Figura 6.9: Áreas para el ejemplo 6.2.

***Solución*:** Véase la figura 6.9 para las áreas específicas.

a) El área en la figura 6.9*a* a la derecha de $z = 1.84$ es igual a 1 menos el área en la tabla A.3 a la izquierda de $z = 1.84$, a saber, $1 - 0.9671 = 0.0329$.

b) El área en la figura 6.9*b* entre $z = -1.97$ y $z = 0.86$ es igual al área a la izquierda de $z = 0.86$ menos el área a la izquierda de $z = -1.97$. A partir de la tabla A.3 encontramos que el área que se desea es $0.8051 - 0.0244 = 0.7807$.

Ejemplo 6.3: Dada una distribución normal estándar, calcule el valor de k tal que

a) $P(Z > k) = 0.3015$, y

b) $P(k < Z < -0.18) = 0.4197$.

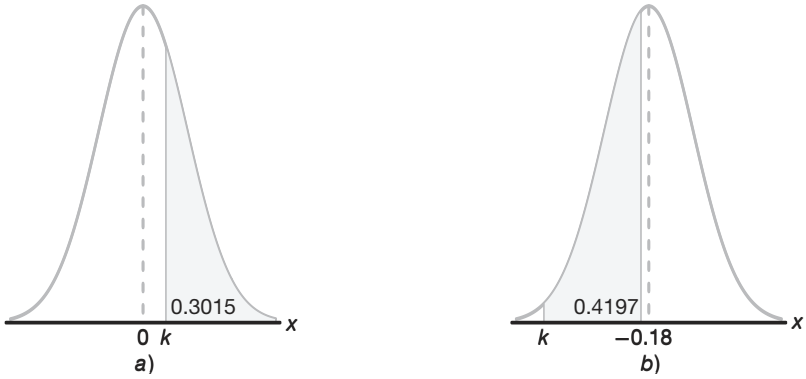


Figura 6.10: Áreas para el ejemplo 6.3.

***Solución*:** La distribución y las áreas deseadas se muestran en la figura 6.10.

a) En la figura 6.10*a* vemos que el valor k que deja un área de 0.3015 a la derecha debe dejar entonces un área de 0.6985 a la izquierda. De la tabla A.3 se sigue que $k = 0.52$.

b) En la tabla A.3 observamos el área total a la izquierda de −0.18 es igual a 0.4286. En la figura 6.10*b* vemos que el área entre k y −0.18 es 0.4197, de manera que el área a la izquierda de k debe ser 0.4286 − 0.4197 = 0.0089. Por lo tanto, a partir de la tabla A.3 tenemos $k = -2.37$.

Ejemplo 6.4: Dada una variable aleatoria X que tiene una distribución normal con $\mu = 50$ y $\sigma = 10$, calcule la probabilidad de que X tome un valor entre 45 y 62.

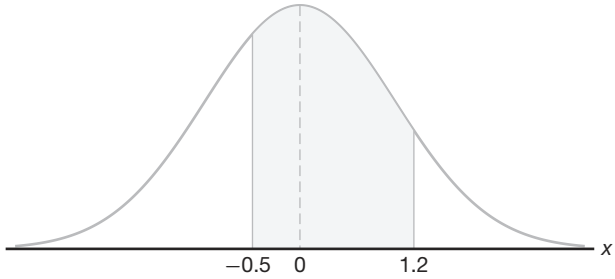


Figura 6.11: Área para el ejemplo 6.4.

***Solución*:** Los valores z que corresponden a $x_1 = 45$ y $x_2 = 62$ son

$$z_1 = \frac{45 - 50}{10} = -0.5 \text{ y } z_2 = \frac{62 - 50}{10} = 1.2.$$

Por lo tanto,

$$P(45 < X < 62) = P(-0.5 < Z < 1.2).$$

$P(-0.5 < Z < 1.2)$ se muestra mediante el área de la región sombreada de la figura 6.11. Esta área se puede calcular restando el área a la izquierda de la ordenada $z = -0.5$ de toda el área a la izquierda de $z = 1.2$. Si usamos la tabla A.3, tenemos

$$\begin{aligned} P(45 < X < 62) &= P(-0.5 < Z < 1.2) = P(Z < 1.2) - P(Z < -0.5) \\ &= 0.8849 - 0.3085 = 0.5764. \end{aligned}$$

Ejemplo 6.5: Dado que X tiene una distribución normal con $\mu = 300$ y $\sigma = 50$, calcule la probabilidad de que X tome un valor mayor que 362.

***Solución*:** La distribución de probabilidad normal que muestra el área sombreada que se desea se presenta en la figura 6.12. Para calcular $P(X > 362)$ necesitamos evaluar el área bajo la curva normal a la derecha de $x = 362$. Esto se puede realizar transformando $x = 362$ al valor z correspondiente, obteniendo el área a la izquierda de z de la tabla A.3 y después restando esta área de 1. Encontramos que

$$z = \frac{362 - 300}{50} = 1.24.$$

De ahí,

$$P(X > 362) = P(Z > 1.24) = 1 - P(Z < 1.24) = 1 - 0.8925 = 0.1075.$$

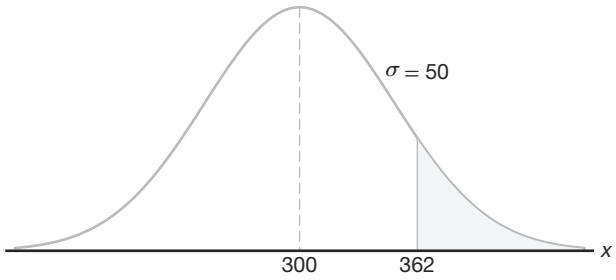


Figura 6.12: Área para el ejemplo 6.5.

De acuerdo con el teorema de Chebyshev en la página 137, la probabilidad de que una variable aleatoria tome un valor dentro de 2 desviaciones estándar de la media es de por lo menos $3/4$. Si la variable aleatoria tiene una distribución normal, los valores z que corresponden a $x_1 = \mu - 2\sigma$ y $x_2 = \mu + 2\sigma$ se calculan fácilmente y son

$$z_1 = \frac{(\mu - 2\sigma) - \mu}{\sigma} = -2 \text{ y } z_2 = \frac{(\mu + 2\sigma) - \mu}{\sigma} = 2.$$

De ahí,

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) = P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 = 0.9544, \end{aligned}$$

que es una afirmación mucho más firme que la que se establece mediante el teorema de Chebyshev.

Uso de la curva normal a la inversa

En ocasiones se nos pide calcular el valor de z que corresponde a una probabilidad específica que cae entre los valores que se listan en la tabla A.3 (véase el ejemplo 6.6). Por conveniencia, siempre elegiremos el valor z que corresponde a la probabilidad tabular que está más cerca de la probabilidad que se especifica.

Los dos ejemplos anteriores se resolvieron al ir primero de un valor de x a un valor z y después calcular el área que se desea. En el ejemplo 6.6 invertimos el proceso y comenzamos con un área o probabilidad conocida, calculamos el valor z y después determinamos x reacomodando la fórmula

$$z = \frac{x - \mu}{\sigma} \text{ para obtener } x = \sigma z + \mu.$$

Ejemplo 6.6: Dada una distribución normal con $\mu = 40$ y $\sigma = 6$, calcule el valor de x que tiene

a) 45% del área a la izquierda, y

b) 14% del área a la derecha.

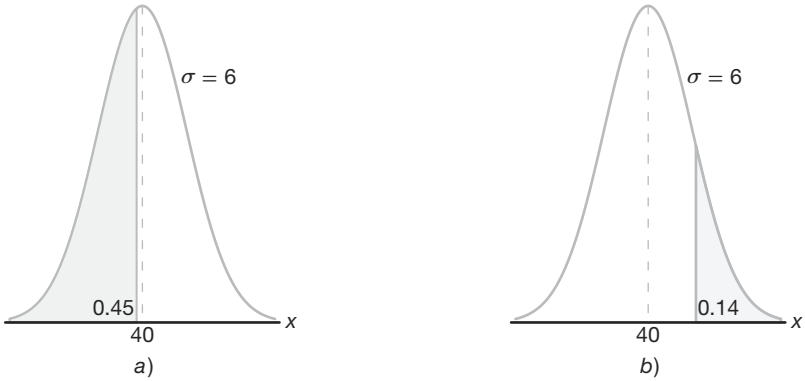


Figura 6.13: Áreas para el ejemplo 6.6.

***Solución*:** *a*) En la figura 6.13*a* se sombrea un área de 0.45 a la izquierda del valor x deseado. Necesitamos un valor z que deje un área de 0.45 a la izquierda. En la tabla A.3 encontramos $P(Z < -0.13) = 0.45$, es decir, que el valor z que se desea es -0.13. Por lo tanto,

$$x = (6)(-0.13) + 40 = 39.22.$$

b) En la figura 6.13*b* sombreamos un área igual a 0.14 a la derecha del valor x deseado. Esta vez necesitamos un valor z que deje 0.14 del área a la derecha y, por lo tanto, un área de 0.86 a la izquierda. De nuevo, a partir de la tabla A.3 encontramos $P(Z < 1.08) = 0.86$, así que el valor z deseado es 1.08 y

$$x = (6)(1.08) + 40 = 46.48.$$

6.4 Aplicaciones de la distribución normal

En los siguientes ejemplos se abordan algunos de los muchos problemas en los que se puede aplicar la distribución normal. El uso de la curva normal para aproximar probabilidades binomiales se estudia en la sección 6.5.

Ejemplo 6.7: Cierto tipo de batería de almacenamiento dura, en promedio, 3.0 años, con una desviación estándar de 0.5 años. Suponga que la duración de la batería se distribuye normalmente y calcule la probabilidad de que una batería determinada dure menos de 2.3 años.

***Solución*:** Empiece construyendo un diagrama como el de la figura 6.14, que muestra la distribución dada de la duración de las baterías y el área deseada. Para calcular la $P(X < 2.3)$ necesitamos evaluar el área bajo la curva normal a la izquierda de 2.3. Esto se logra calculando el área a la izquierda del valor z correspondiente. De donde encontramos que

$$z = \frac{2.3 - 3}{0.5} = -1.4,$$

y entonces, usando la tabla A.3, tenemos

$$P(X < 2.3) = P(Z < -1.4) = 0.0808.$$

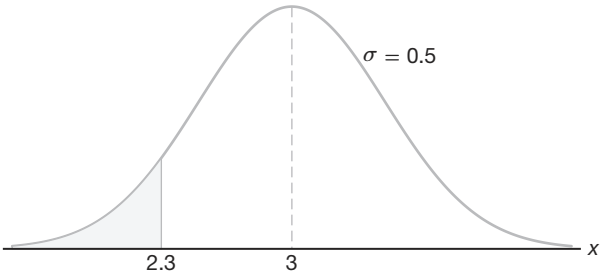


Figura 6.14: Área para el ejemplo 6.7.

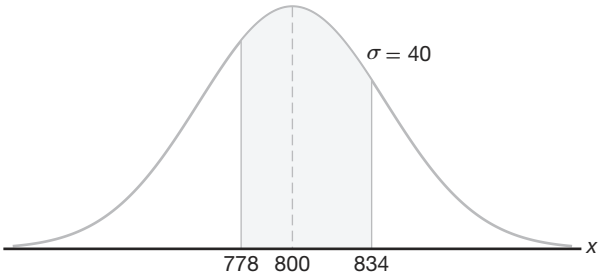


Figura 6.15: Área para el ejemplo 6.8.

Ejemplo 6.8 Una empresa de material eléctrico fabrica bombillas de luz cuya duración, antes de quemarse, se distribuye normalmente con una media igual a 800 horas y una desviación estándar de 40 horas. Calcule la probabilidad de que una bombilla se queme entre 778 y 834 horas.

***Solución*:** La distribución de vida de las bombillas se ilustra en la figura 6.15. Los valores z que corresponden a $x_1 = 778$ y $x_2 = 834$ son

$$z_1 = \frac{778 - 800}{40} = -0.55 \text{ y } z_2 = \frac{834 - 800}{40} = 0.85.$$

Por lo tanto,

$$\begin{aligned} P(778 < X < 834) &= P(-0.55 < Z < 0.85) = P(Z < 0.85) - P(Z < -0.55) \\ &= 0.8023 - 0.2912 = 0.5111. \end{aligned}$$

Ejemplo 6.9: En un proceso industrial el diámetro de un cojinete de bolas es una medida importante. El comprador establece que las especificaciones en el diámetro sean 3.0 ± 0.01 cm. Esto

implica que no se aceptará ninguna parte que no cumpla estas especificaciones. Se sabe que en el proceso el diámetro de un cojinete tiene una distribución normal con media $\mu = 3.0$ y una desviación estándar $\sigma = 0.005$. En promedio, ¿cuántos de los cojinetes fabricados se descartarán?

***Solución*:** La distribución de los diámetros se ilustra en la figura 6.16. Los valores que corresponden a los límites especificados son $x_1 = 2.99$ y $x_2 = 3.01$. Los valores z correspondientes son

$$z_1 = \frac{2.99 - 3.0}{0.005} = -2.0 \text{ y } z_2 = \frac{3.01 - 3.0}{0.005} = +2.0.$$

Por lo tanto,

$$P(2.99 < X < 3.01) = P(-2.0 < Z < 2.0).$$

A partir de la tabla A.3, $P(Z < -2.0) = 0.0228$. Debido a la simetría de la distribución normal, encontramos que

$$P(Z < -2.0) + P(Z > 2.0) = 2(0.0228) = 0.0456.$$

Como resultado se anticipa que, en promedio, se descartarán 4.56% de los cojinetes fabricados.

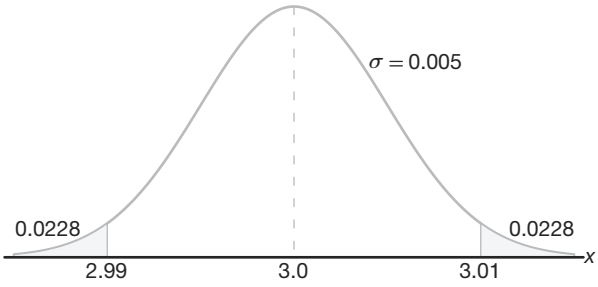


Figura 6.16: Área para el ejemplo 6.9.

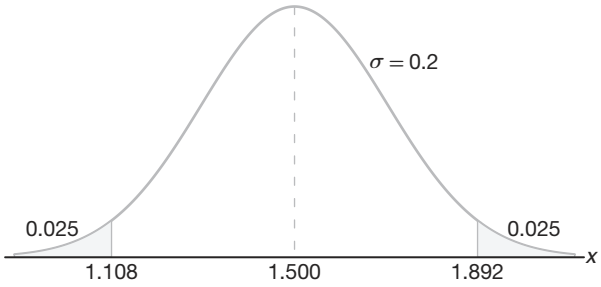


Figura 6.17: Especificaciones para el ejemplo 6.10.

Ejemplo 6.10: Se utilizan medidores para rechazar todos los componentes en los que cierta dimensión no esté dentro de la especificación $1.50 \pm d$. Se sabe que esta medida se distribuye normalmente con una media de 1.50 y una desviación estándar de 0.2. Determine el valor d tal que las especificaciones "cubran" 95% de las mediciones.

***Solución*:** A partir de la tabla A.3 sabemos que

$$P(-1.96 < Z < 1.96) = 0.95.$$

Por lo tanto,

$$1.96 = \frac{(1.50 + d) - 1.50}{0.2},$$

de la que obtenemos

$$d = (0.2)(1.96) = 0.392.$$

En la figura 6.17 se muestra una ilustración de las especificaciones.

Ejemplo 6.11: Cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms. Si se supone que la resistencia sigue una distribución normal y que se puede medir con cualquier grado de precisión, ¿qué porcentaje de resistencias tendrán una resistencia que exceda 43 ohms?

***Solución*:** Se obtiene un porcentaje multiplicando la frecuencia relativa por 100%. Como la frecuencia relativa para un intervalo es igual a la probabilidad de caer en el intervalo, debemos calcular el área a la derecha de $x = 43$ en la figura 6.18. Esto se puede hacer transformando $x = 43$ al valor z correspondiente, con lo cual se obtiene el área a la izquierda de z de la tabla A.3, y después se resta esta área de 1. Encontramos que

$$z = \frac{43 - 40}{2} = 1.5.$$

Por lo tanto,

$$P(X > 43) = P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668.$$

Así, 6.68% de las resistencias tendrán una resistencia que exceda 43 ohms. ∎

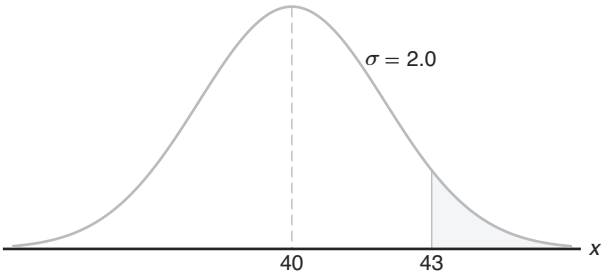


Figura 6.18: Área para el ejemplo 6.11.

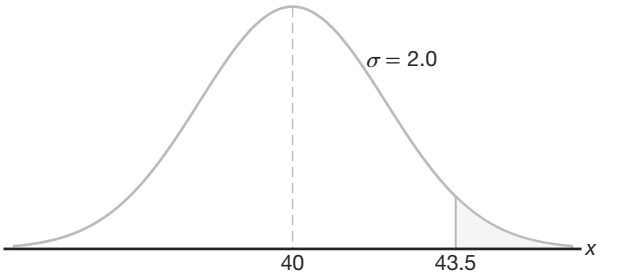


Figura 6.19: Área para el ejemplo 6.12.

Ejemplo 6.12: Calcule el porcentaje de resistencias que excedan 43 ohms para el ejemplo 6.11 si la resistencia se mide al ohm más cercano.

***Solución*:** Este problema difiere del ejemplo 6.11 en que ahora asignamos una medida de 43 ohms a todos los resistores cuyas resistencias sean mayores que 42.5 y menores que 43.5. Lo que estamos haciendo realmente es aproximar una distribución discreta por medio de una distribución continua normal. El área que se requiere es la región sombreada a la derecha de 43.5 en la figura 6.19. Encontramos ahora que

$$z = \frac{43.5 - 40}{2} = 1.75.$$

En consecuencia,

$$P(X > 43.5) = P(Z > 1.75) = 1 - P(Z < 1.75) = 1 - 0.9599 = 0.0401.$$

Por lo tanto, 4.01% de las resistencias exceden 43 ohms cuando se miden al ohm más cercano. La diferencia 6.68% – 4.01% = 2.67% entre esta respuesta y la del ejemplo 6.11 representa todos los valores de resistencias mayores que 43 y menores que 43.5, que ahora se registran como de 43 ohms. ∎

Ejemplo 6.13: La calificación promedio para un examen es 74 y la desviación estándar es 7. Si 12% del grupo obtiene *A* y las calificaciones siguen una curva que tiene una distribución normal, ¿cuál es la *A* más baja posible y la *B* más alta posible?

***Solución*:** En este ejemplo comenzamos con un área de probabilidad conocida, calculamos el valor z y después determinamos x con la fórmula $x = \sigma z + \mu$. Un área de 0.12, que corresponde a la fracción de estudiantes que reciben *A,* está sombreada en la figura 6.20. Necesitamos un valor z que deje 0.12 del área a la derecha y, por lo tanto, un área de 0.88 a la izquierda. A partir de la tabla A.3, $P(Z < 1.18)$ tiene el valor más cercano a 0.88, de manera que el valor z que se desea es 1.18. En consecuencia,

$$x = (7)(1.18) + 74 = 82.26.$$

Por lo tanto, la *A* más baja es 83 y la *B* más alta es 82.

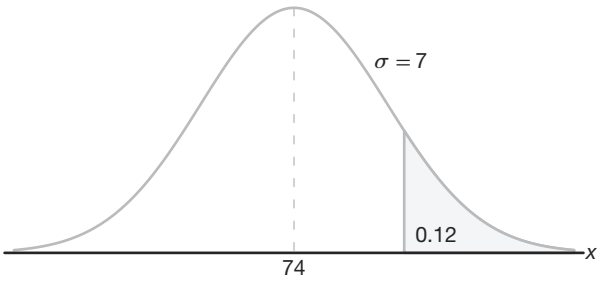


Figura 6.20: Área para el ejemplo 6.13.

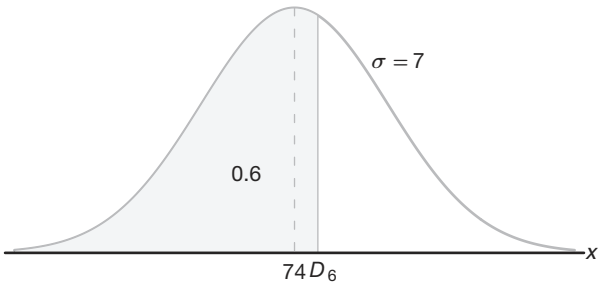


Figura 6.21: Área para el ejemplo 6.14.

Ejemplo 6.14: Remítase al ejemplo 6.13 y calcule el sexto decil.

***Solución*:** El sexto decil, escrito como D_6, es el valor x que deja 60% del área a la izquierda, como se muestra en la figura 6.21. En la tabla A.3 encontramos que $P(Z < 0.25) \approx 0.6$, de manera que el valor z deseado es 0.25. Ahora, $x = (7)\,(0.25) + 74 = 75.75$. Por lo tanto, $D_6 = 75.75$. Es decir, 60% de las calificaciones son 75 o menos.

Ejercicios

6.1 Dada una distribución continua uniforme, demuestre que

a) $\mu = \frac{A+B}{2}$, y

b) $\sigma^2 = \frac{(B-A)^2}{12}$.

6.2 Suponga que X tiene una distribución continua uniforme de 1 a 5. Determine la probabilidad condicional $P(X > 2.5 \mid X \leq 4)$.

6.3 La cantidad de café diaria, en litros, que sirve una máquina que se localiza en el vestíbulo de un aeropuerto es una variable aleatoria X que tiene una distribución continua uniforme con $A = 7$ y $B = 10$. Calcule la probabilidad de que en un día determinado la cantidad de café que sirve esta máquina sea

a) a lo sumo 8.8 litros;

b) más de 7.4 litros, pero menos de 9.5 litros;

c) al menos 8.5 litros.

6.4 Un autobús llega cada 10 minutos a una parada. Se supone que el tiempo de espera para un individuo en particular es una variable aleatoria con distribución continua uniforme.

a) ¿Cuál es la probabilidad de que el individuo espere más de 7 minutos?
b) ¿Cuál es la probabilidad de que el individuo espere entre 2 y 7 minutos?

6.5 Dada una distribución normal estándar, calcule el área bajo la curva que está
a) a la izquierda de $z = -1.39$;
b) a la derecha de $z = 1.96$;
c) entre $z = -2.16$ y $z = -0.65$;
d) a la izquierda de $z = 1.43$;
e) a la derecha de $z = -0.89$;
f) entre $z = -0.48$ y $z = 1.74$.

6.6 Calcule el valor de z si el área bajo una curva normal estándar
a) a la derecha de z es 0.3622;
b) a la izquierda de z es 0.1131;
c) entre 0 y z, con $z > 0$, es 0.4838;
d) entre $-z$ y z, con $z > 0$, es 0.9500.

6.7 Dada una distribución normal estándar, calcule el valor de k tal que
a) $P(Z > k) = 0.2946$;
b) $P(Z < k) = 0.0427$;
c) $P(-0.93 < Z < k) = 0.7235$.

6.8 Dada una distribución normal con $\mu = 30$ y $\sigma = 6$, calcule
a) el área de la curva normal a la derecha de $x = 17$;
b) el área de la curva normal a la izquierda de $x = 22$;
c) el área de la curva normal entre $x = 32$ y $x = 41$;
d) el valor de x que tiene 80% del área de la curva normal a la izquierda;
e) los dos valores de x que contienen 75% central del área de la curva normal.

6.9 Dada la variable X normalmente distribuida con una media de 18 y una desviación estándar de 2.5, calcule
a) $P(X < 15)$;
b) el valor de k tal que $P(X < k) = 0.2236$;
c) el valor de k tal que $P(X > k) = 0.1814$;
d) $P(17 < X < 21)$.

6.10 De acuerdo con el teorema de Chebyshev, la probabilidad de que cualquier variable aleatoria tome un valor dentro de 3 desviaciones estándar de la media es de al menos $8/9$. Si se sabe que la distribución de probabilidad de una variable aleatoria X es normal con media μ y varianza σ^2, ¿cuál es el valor exacto de $P(\mu - 3\sigma < X < \mu + 3\sigma)$?

6.11 Una máquina expendedora de bebidas gaseosas se regula para que sirva un promedio de 200 mililitros por vaso. Si la cantidad de bebida se distribuye normalmente con una desviación estándar igual a 15 mililitros,
a) ¿qué fracción de los vasos contendrá más de 224 mililitros?
b) ¿cuál es la probabilidad de que un vaso contenga entre 191 y 209 mililitros?
c) ¿cuántos vasos probablemente se derramarán si se utilizan vasos de 230 mililitros para las siguientes 1000 bebidas?
d) ¿por debajo de qué valor obtendremos el 25% más bajo en el llenado de las bebidas?

6.12 Las barras de pan de centeno que cierta panadería distribuye a las tiendas locales tienen una longitud promedio de 30 centímetros y una desviación estándar de 2 centímetros. Si se supone que las longitudes están distribuidas normalmente, ¿qué porcentaje de las barras son
a) más largas que 31.7 centímetros?
b) de entre 29.3 y 33.5 centímetros de longitud?
c) más cortas que 25.5 centímetros?

6.13 Un investigador informa que unos ratones a los que primero se les restringen drásticamente sus dietas y después se les enriquecen con vitaminas y proteínas vivirán un promedio de 40 meses. Si suponemos que la vida de tales ratones se distribuye normalmente, con una desviación estándar de 6.3 meses, calcule la probabilidad de que un ratón determinado viva
a) más de 32 meses;
b) menos de 28 meses;
c) entre 37 y 49 meses.

6.14 El diámetro interior del anillo de un pistón terminado se distribuye normalmente con una media de 10 centímetros y una desviación estándar de 0.03 centímetros.
a) ¿Qué proporción de anillos tendrá diámetros interiores que excedan 10.075 centímetros?
b) ¿Cuál es la probabilidad de que el anillo de un pistón tenga un diámetro interior de entre 9.97 y 10.03 centímetros?
c) ¿Por debajo de qué valor del diámetro interior caerá el 15% de los anillos de pistón?

6.15 Un abogado viaja todos los días de su casa en los suburbios a su oficina en el centro de la ciudad. El tiempo promedio para un viaje sólo de ida es de 24 minutos, con una desviación estándar de 3.8 minutos. Si se supone que la distribución de los tiempos de viaje está distribuida normalmente.
a) ¿Cuál es la probabilidad de que un viaje tome al menos $1/2$ hora?
b) Si la oficina abre a las 9:00 A.M. y él sale diario de su casa a las 8:45 A.M., ¿qué porcentaje de las veces llegará tarde al trabajo?

c) Si sale de su casa a las 8:35 A.M. y el café se sirve en la oficina de 8:50 A.M. a 9:00 A.M., ¿cuál es la probabilidad de que se pierda el café?
d) Calcule la duración mayor en la que se encuentra el 15% de los viajes más lentos.
e) Calcule la probabilidad de que 2 de los siguientes 3 viajes tomen al menos $1/2$ hora.

6.16 En el ejemplar de noviembre de 1990 de *Chemical Engineering Progress*, un estudio analiza el porcentaje de pureza del oxígeno de cierto proveedor. Suponga que la media fue de 99.61, con una desviación estándar de 0.08. Suponga que la distribución del porcentaje de pureza fue aproximadamente normal.
a) ¿Qué porcentaje de los valores de pureza esperaría que estuvieran entre 99.5 y 99.7?
b) ¿Qué valor de pureza esperaría que excediera exactamente 5% de la población?

6.17 La vida promedio de cierto tipo de motor pequeño es de 10 años, con una desviación estándar de 2 años. El fabricante reemplaza gratis todos los motores que fallen dentro del periodo de garantía. Si estuviera dispuesto a reemplazar sólo 3% de los motores que fallan, ¿cuánto tiempo de garantía debería ofrecer? Suponga que la duración de un motor sigue una distribución normal.

6.18 La estatura de 1000 estudiantes se distribuye normalmente con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se supone que las estaturas se redondean al medio centímetro más cercano, ¿cuántos de estos estudiantes esperaría que tuvieran una estatura
a) menor que 160.0 centímetros?
b) de entre 171.5 y 182.0 centímetros inclusive?
c) igual a 175.0 centímetros?
d) mayor o igual que 188.0 centímetros?

6.19 Una empresa paga a sus empleados un salario promedio de \$15.90 por hora, con una desviación estándar de \$1.50. Si los salarios se distribuyen aproximadamente de forma normal y se redondean al centavo más cercano,
a) ¿qué porcentaje de los trabajadores recibe salarios de entre \$13.75 y \$16.22 por hora?
b) ¿el 5% de los salarios más altos por hora de los empleados es mayor a qué cantidad?

6.20 Los pesos de un gran número de *poodle* miniatura se distribuyen aproximadamente de forma normal con una media de 8 kilogramos y una desviación estándar de 0.9 kilogramos. Si las mediciones se redondean al décimo de kilogramo más cercano, calcule la fracción de estos *poodle* con pesos
a) por arriba de 9.5 kilogramos;
b) a lo sumo 8.6 kilogramos;
c) entre 7.3 y 9.1 kilogramos.

6.21 La resistencia a la tensión de cierto componente de metal se distribuye normalmente con una media de 10,000 kilogramos por centímetro cuadrado y una desviación estándar de 100 kilogramos por centímetro cuadrado. Las mediciones se redondean a los 50 kilogramos por centímetro cuadrado más cercanos.
a) ¿Qué proporción de estos componentes excede a 10,150 kilogramos por centímetro cuadrado de resistencia a la tensión?
b) Si las especificaciones requieren que todos los componentes tengan una resistencia a la tensión de entre 9800 y 10,200 kilogramos por centímetro cuadrado, ¿qué proporción de piezas esperaría que se descartara?

6.22 Si un conjunto de observaciones se distribuye de manera normal, ¿qué porcentaje de éstas difieren de la media en
a) más de 1.3σ?
b) menos de 0.52σ?

6.23 El coeficiente intelectual (CI) de 600 aspirantes a cierta universidad se distribuye aproximadamente de forma normal con una media de 115 y una desviación estándar de 12. Si la universidad requiere un CI de al menos 95, ¿cuántos de estos estudiantes serán rechazados con base en éste sin importar sus otras calificaciones? Tome en cuenta que el CI de los aspirantes se redondea al entero más cercano.

6.5 Aproximación normal a la binomial

Las probabilidades asociadas con experimentos binomiales se obtienen fácilmente a partir de la fórmula $b(x; n, p)$ de la distribución binomial o de la tabla A.1 cuando n es pequeña. Además, las probabilidades binomiales están disponibles en muchos paquetes de software. Sin embargo, resulta aleccionador conocer la relación entre la distribución binomial y la normal. En la sección 5.5 explicamos cómo se puede utilizar la distribución de Poisson para aproximar probabilidades binomiales cuando n es muy grande y p se acerca mucho a 0 o a 1. Tanto la distribución binomial como la de Poisson son

discretas. La primera aplicación de una distribución continua de probabilidad para aproximar probabilidades sobre un espacio muestral discreto se demostró en el ejemplo 6.12, donde se utilizó la curva normal. La distribución normal a menudo es una buena aproximación a una distribución discreta cuando la última adquiere una forma de campana simétrica. Desde un punto de vista teórico, algunas distribuciones convergen a la normal a medida que sus parámetros se aproximan a ciertos límites. La distribución normal es una distribución de aproximación conveniente, ya que la función de distribución acumulativa se tabula con mucha facilidad. La distribución binomial se aproxima bien por medio de la normal en problemas prácticos cuando se trabaja con la función de distribución acumulativa. Ahora plantearemos un teorema que nos permitirá utilizar áreas bajo la curva normal para aproximar propiedades binomiales cuando n es suficientemente grande.

Teorema 6.3: Si X es una variable aleatoria binomial con media $\mu = np$ y varianza $\sigma^2 = npq$, entonces la forma limitante de la distribución de

$$Z = \frac{X - np}{\sqrt{npq}},$$

conforme $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

Resulta que la distribución normal con $\mu = np$ y $\sigma^2 = np(1-p)$ no sólo ofrece una aproximación muy precisa a la distribución binomial cuando n es grande y p no está extremadamente cerca de 0 o de 1, sino que también brinda una aproximación bastante buena aun cuando n es pequeña y p está razonablemente cerca de $1/2$.

Para ilustrar la aproximación normal a la distribución binomial primero dibujamos el histograma para $b(x; 15, 0.4)$ y después superponemos la curva normal particular con la misma media y varianza que la variable binomial X. En consecuencia, dibujamos una curva normal con

$$\mu = np = (15)(0.4) = 6 \text{ y } \sigma^2 = npq = (15)(0.4)(0.6) = 3.6.$$

El histograma de $b(x; 15, 0.4)$ y la curva normal superpuesta correspondiente, que está determinada por completo por su media y su varianza, se ilustran en la figura 6.22.

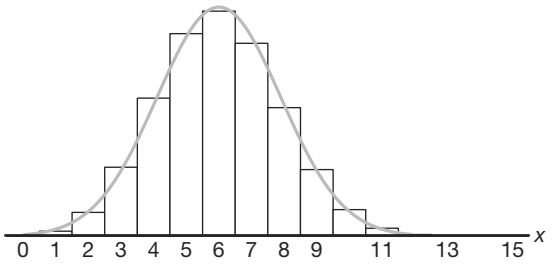


Figura 6.22: Aproximación normal de $b(x; 15, 0.4)$.

La probabilidad exacta de que la variable aleatoria binomial X tome un valor determinado x es igual al área de la barra cuya base se centra en x. Por ejemplo, la probabilidad exacta de que X tome el valor 4 es igual al área del rectángulo con base centrada en $x = 4$. Si usamos la tabla A.1, encontramos que esta área es

$$P(X = 4) = b(4; 15, 0.4) = 0.1268,$$

que es aproximadamente igual al área de la región sombreada bajo la curva normal entre las dos ordenadas $x_1 = 3.5$ y $x_2 = 4.5$ en la figura 6.23. Al convertir a valores z, tenemos

$$z_1 = \frac{3.5 - 6}{1.897} = -1.32 \quad \text{y} \quad z_2 = \frac{4.5 - 6}{1.897} = -0.79.$$

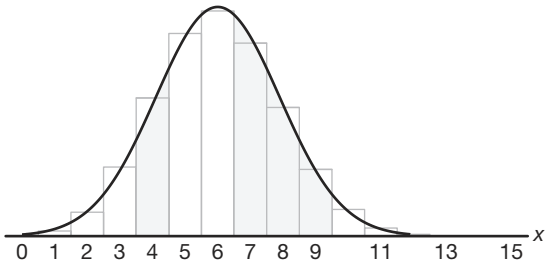


Figura 6.23: Aproximación normal de $b(x; 15, 0.4)$ y $\sum_{x=7}^{9} b(x; 15, 0.4)$.

Si X es una variable aleatoria binomial y Z una variable normal estándar, entonces,

$$\begin{aligned} P(X = 4) &= b(4; 15, 0.4) \approx P(-1.32 < Z < -0.79) \\ &= P(Z < -0.79) - P(Z < -1.32) = 0.2148 - 0.0934 = 0.1214. \end{aligned}$$

Esto se aproxima bastante al valor exacto de 0.1268.

La aproximación normal es más útil en el cálculo de sumatorias binomiales para valores grandes de n. Si nos remitimos a la figura 6.23, nos podríamos interesar en la probabilidad de que X tome un valor de 7 a 9. La probabilidad exacta es dada por

$$\begin{aligned} P(7 \le X \le 9) &= \sum_{x=0}^{9} b(x; 15, 0.4) - \sum_{x=0}^{6} b(x; 15, 0.4) \\ &= 0.9662 - 0.6098 = 0.3564, \end{aligned}$$

que es igual a la sumatoria de las áreas de los rectángulos cuyas bases están centradas en $x = 7$, 8 y 9. Para la aproximación normal calculamos el área de la región sombreada bajo la curva entre las ordenadas $x_1 = 6.5$ y $x_2 = 9.5$ de la figura 6.23. Los valores z correspondientes son

$$z_1 = \frac{6.5 - 6}{1.897} = 0.26 \quad \text{y} \quad z_2 = \frac{9.5 - 6}{1.897} = 1.85.$$

Ahora,

$$P(7 \leq X \leq 9) \approx P(0.26 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.26)$$
$$= 0.9678 - 0.6026 = 0.3652.$$

Una vez más, la aproximación de la curva normal ofrece un valor que se acerca al valor exacto de 0.3564. El grado de exactitud, que depende de qué tan bien se ajuste la curva al histograma, se incrementa a medida que aumenta n. Esto es particularmente cierto cuando p no está muy cerca de $1/2$ y el histograma ya no es simétrico. Las figuras 6.24 y 6.25 muestran los histogramas para $b(x; 6, 0.2)$ y $b(x; 15, 0.2)$, respectivamente. Es evidente que una curva normal se ajustará mucho mejor al histograma cuando $n = 15$ que cuando $n = 6$.

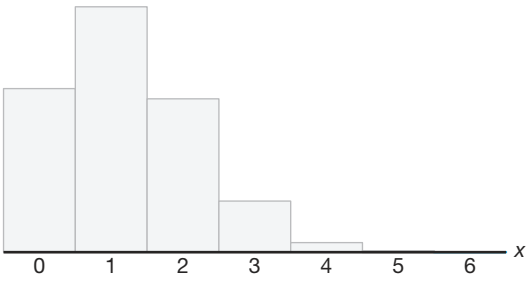


Figura 6.24: Histograma para $b(x; 6, 0.2)$.

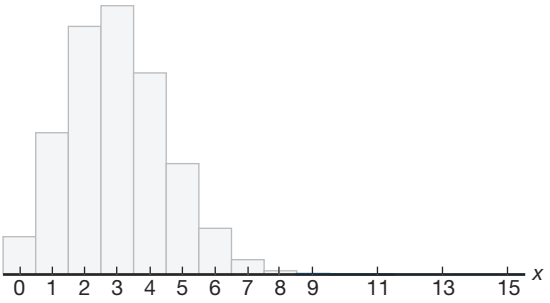


Figura 6.25: Histograma para $b(x; 15, 0.2)$.

En las ilustraciones de la aproximación normal a la binomial se hizo evidente que si buscamos el área bajo la curva normal hacia la izquierda de, digamos x, es más preciso utilizar $x + 0.5$. Esto es una corrección para dar cabida al hecho de que una distribución discreta se aproxima mediante una distribución continua. La corrección $+\mathbf{0.5}$ se llama **corrección de continuidad**. La explicación anterior conduce a la siguiente aproximación normal formal a la binomial.

Aproximación normal a la distribución binomial

Sea X una variable aleatoria binomial con parámetros n y p. Para una n grande, X tiene aproximadamente una distribución normal con $\mu = np$ y $\sigma^2 = npq = np(1-p)$ y

$$P(X \leq x) = \sum_{k=0}^{x} b(k; n, p)$$
$$\approx \text{área bajo la curva normal a la izquierda de } x + 0.5$$
$$= P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{npq}}\right),$$

y la aproximación será buena si np y $n(1-p)$ son mayores que o iguales a 5.

Como indicamos antes, la calidad de la aproximación es muy buena para n grande. Si p está cerca de $1/2$, un tamaño de la muestra moderado o pequeño será suficiente para una aproximación razonable. Ofrecemos la tabla 6.1 como una indicación de la calidad

de la aproximación. Se presentan tanto la aproximación normal como las probabilidades binomiales acumulativas reales. Observe que en $p = 0.05$ y $p = 0.10$ la aproximación es muy burda para $n = 10$. Sin embargo, incluso para $n = 10$, observe la mejoría para $p = 0.50$. Por otro lado, cuando p es fija en $p = 0.05$, observe cómo mejora la aproximación conforme vamos de $n = 20$ a $n = 100$.

Tabla 6.1: Aproximación normal y probabilidades binomiales acumulativas reales

	$p = 0.05, n = 10$		$p = 0.10, n = 10$		$p = 0.50, n = 10$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.5987	0.5000	0.3487	0.2981	0.0010	0.0022
1	0.9139	0.9265	0.7361	0.7019	0.0107	0.0136
2	0.9885	0.9981	0.9298	0.9429	0.0547	0.0571
3	0.9990	1.0000	0.9872	0.9959	0.1719	0.1711
4	1.0000	1.0000	0.9984	0.9999	0.3770	0.3745
5			1.0000	1.0000	0.6230	0.6255
6					0.8281	0.8289
7					0.9453	0.9429
8					0.9893	0.9864
9					0.9990	0.9978
10					1.0000	0.9997

	$p = 0.05$					
	$n = 20$		$n = 50$		$n = 100$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.3585	0.3015	0.0769	0.0968	0.0059	0.0197
1	0.7358	0.6985	0.2794	0.2578	0.0371	0.0537
2	0.9245	0.9382	0.5405	0.5000	0.1183	0.1251
3	0.9841	0.9948	0.7604	0.7422	0.2578	0.2451
4	0.9974	0.9998	0.8964	0.9032	0.4360	0.4090
5	0.9997	1.0000	0.9622	0.9744	0.6160	0.5910
6	1.0000	1.0000	0.9882	0.9953	0.7660	0.7549
7			0.9968	0.9994	0.8720	0.8749
8			0.9992	0.9999	0.9369	0.9463
9			0.9998	1.0000	0.9718	0.9803
10			1.0000	1.0000	0.9885	0.9941

Ejemplo 6.15: Un paciente que padece una rara enfermedad de la sangre tiene 0.4 de probabilidad de recuperarse. Si se sabe que 100 personas contrajeron esta enfermedad, ¿cuál es la probabilidad de que sobrevivan menos de 30?

***Solución*:** Representemos con la variable binomial X el número de pacientes que sobreviven. Como $n = 100$, deberíamos obtener resultados muy precisos usando la aproximación de la curva normal con

$$\mu = np = (100)(0.4) = 40 \text{ y } \sigma = \sqrt{npq} = \sqrt{(100)(0.4)(0.6)} = 4.899.$$

Para obtener la probabilidad que se desea, tenemos que calcular el área a la izquierda de $x = 29.5$.

El valor z que corresponde a 29.5 es

$$z = \frac{29.5 - 40}{4.899} = -2.14,$$

y la probabilidad de que menos de 30 de los 100 pacientes sobrevivan está dada por la región sombreada en la figura 6.26. Por lo tanto,

$$P(X < 30) \approx P(Z < -2.14) = 0.0162.$$

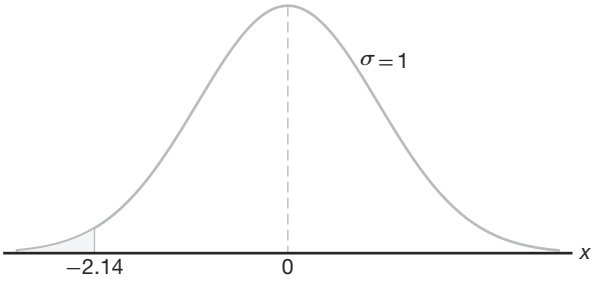


Figura 6.26: Área para el ejemplo 6.15.

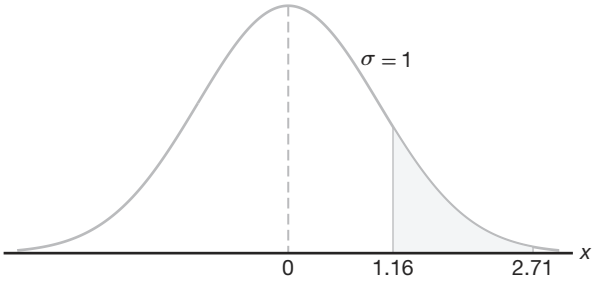


Figura 6.27: Área para el ejemplo 6.16.

Ejemplo 6.16: Un examen de opción múltiple tiene 200 preguntas, cada una con 4 respuestas posibles, de las que sólo una es la correcta. ¿Cuál es la probabilidad de que solamente adivinando se obtengan de 25 a 30 respuestas correctas para 80 de los 200 problemas sobre los que el estudiante no tiene conocimientos?

***Solución*:** La probabilidad de adivinar una respuesta correcta para cada una de las 80 preguntas es $p = 1/4$. Si X representa el número de respuestas correctas sólo porque se adivinaron, entonces,

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 1/4).$$

Al usar la aproximación de la curva normal con

$$\mu = np = (80)\left(\frac{1}{4}\right) = 20$$

y

$$\sigma = \sqrt{npq} = \sqrt{(80)(1/4)(3/4)} = 3.873,$$

necesitamos el área entre $x_1 = 24.5$ y $x_2 = 30.5$. Los valores z correspondientes son

$$z_1 = \frac{24.5 - 20}{3.873} = 1.16 \text{ y } z_2 = \frac{30.5 - 20}{3.873} = 2.71.$$

La probabilidad de adivinar correctamente de 25 a 30 preguntas es dada por la región sombreada de la figura 6.27. En la tabla A.3 encontramos que

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 0.25) \approx P(1.16 < Z < 2.71)$$
$$= P(Z < 2.71) - P(Z < 1.16) = 0.9966 - 0.8770 = 0.1196.$$

Ejercicios

6.24 Se lanza una moneda 400 veces. Utilice la aproximación a la curva normal para calcular la probabilidad de obtener
a) entre 185 y 210 caras;
b) exactamente 205 caras;
c) menos de 176 o más de 227 caras.

6.25 En un proceso para fabricar un componente electrónico, 1% de los artículos resultan defectuosos. Un plan de control de calidad consiste en seleccionar 100 artículos de un proceso de producción y detenerlo o continuar con él si ninguno está defectuoso. Use la aproximación normal a la binomial para calcular
a) la probabilidad de que el proceso continúe con el plan de muestreo descrito;
b) la probabilidad de que el proceso continúe aun si éste va mal (es decir, si la frecuencia de componentes defectuosos cambió a 5.0% de defectuosos).

6.26 Un proceso produce 10% de artículos defectuosos. Si se seleccionan al azar 100 artículos del proceso, ¿cuál es la probabilidad de que el número de defectuosos
a) exceda los 13?
b) sea menor que 8?

6.27 Un paciente tiene 0.9 de probabilidad de recuperarse de una operación de corazón delicada. De los siguientes 100 pacientes que se someten a esta operación, ¿cuál es la probabilidad de que
a) sobrevivan entre 84 y 95 inclusive?
b) sobrevivan menos de 86?

6.28 Investigadores de la Universidad George Washington y del Instituto Nacional de Salud informan que aproximadamente 75% de las personas cree que "los tranquilizantes funcionan muy bien para lograr que una persona esté más tranquila y relajada". De las siguientes 80 personas entrevistadas, ¿cuál es la probabilidad de que
a) al menos 50 tengan esta opinión?
b) a lo sumo 56 tengan esta opinión?

6.29 Si 20% de los residentes de una ciudad de Estados Unidos prefieren un teléfono blanco sobre cualquier otro color disponible, ¿cuál es la probabilidad de que, de los siguientes 1000 teléfonos que se instalen en esa ciudad,
a) entre 170 y 185 sean blancos?
b) al menos 210 pero no más de 225 sean blancos?

6.30 Un fabricante de medicamentos sostiene que cierto medicamento cura una enfermedad de la sangre, en promedio, 80% de las veces. Para verificar la aseveración, inspectores gubernamentales utilizan el medicamento en una muestra de 100 individuos y deciden aceptar la afirmación si se curan 75 o más.
a) ¿Cuál es la probabilidad de que los inspectores gubernamentales rechacen la aseveración si la probabilidad de curación es, de hecho, de 0.8?
b) ¿Cuál es la probabilidad de que el gobierno acepte la afirmación si la probabilidad de curación resulta tan baja como 0.7?

6.31 Una sexta parte de los estudiantes de primer año que entran a una escuela estatal grande provienen de otros estados. Si son asignados al azar a los 180 dormitorios de un edificio, ¿cuál es la probabilidad de que en un determinado dormitorio al menos una quinta parte de los estudiantes provenga de otro estado?

6.32 Una empresa farmacéutica sabe que aproximadamente 5% de sus píldoras anticonceptivas no contiene la cantidad suficiente de un ingrediente, lo que las vuelve ineficaces. ¿Cuál es la probabilidad de que menos de 10 píldoras en una muestra de 200 sean ineficaces?

6.33 Estadísticas publicadas por la National Highway Traffic Safety Administration y el National Safety Council revelan que en una noche promedio de fin de semana, uno de cada 10 conductores está ebrio. Si la siguiente noche de sábado se revisan 400 conductores al azar, ¿cuál es la probabilidad de que el número de conductores ebrios sea
a) menor que 32?
b) mayor que 49?
c) al menos 35 pero menos que 47?

6.34 Un par de dados se lanza 180 veces. ¿Cuál es la probabilidad de que ocurra un total de 7
a) al menos 25 veces?
b) entre 33 y 41 veces?
c) exactamente 30 veces?

6.35 Una empresa produce partes componentes para un motor. Las especificaciones de las partes sugieren que sólo 95% de los artículos las cumplen. Las partes para los clientes se embarcan en lotes de 100.
a) ¿Cuál es la probabilidad de que más de 2 artículos estén defectuosos en un lote determinado?
b) ¿Cuál es la probabilidad de que más de 10 artículos de un lote estén defectuosos?

6.36 Una práctica común por parte de las aerolíneas consiste en vender más boletos que el número real de asientos para un vuelo específico porque los clientes que compran boletos no siempre se presentan a abordar el avión. Suponga que el porcentaje de pasajeros que no se presentan a la hora del vuelo es de 2%. Para un vuelo particular con 197 asientos, se vendieron un total

de 200 boletos. ¿Cuál es la probabilidad de que la aerolínea haya sobrevendido el vuelo?

6.37 El nivel X de colesterol en la sangre en muchachos de 14 años tiene aproximadamente una distribución normal, con una media de 170 y una desviación estándar de 30.

a) Determine la probabilidad de que el nivel de colesterol en la sangre de un muchacho de 14 años elegido al azar exceda 230.

b) En una escuela secundaria hay 300 muchachos de 14 años. Determine la probabilidad de que por lo menos 8 de ellos tengan un nivel de colesterol superior a 230.

6.38 Una empresa de telemarketing tiene una máquina especial para abrir cartas que abre y extrae el contenido de los sobres. Si un sobre se colocara de forma incorrecta en la máquina, no se podría extraer su contenido, o incluso se podría dañar. En este caso se dice que "falló" la máquina.

a) Si la probabilidad de que falle la máquina es de 0.01, ¿cuál es la probabilidad de que ocurra más de una falla en un lote de 20 sobres?

b) Si la probabilidad de que falle la máquina es de 0.01 y se abrirá un lote de 500 sobres, ¿cuál es la probabilidad de que ocurran más de 8 fallas?

6.6 Distribución gamma y distribución exponencial

Aunque la distribución normal se puede utilizar para resolver muchos problemas de ingeniería y ciencias, aún hay numerosas situaciones que requieren diferentes tipos de funciones de densidad. En esta sección se estudiarán dos de estas funciones de densidad, la **distribución gamma** y la **distribución exponencial**.

Resulta que la distribución exponencial es un caso especial de la distribución gamma, y ambas tienen un gran número de aplicaciones. La distribución exponencial y la distribución gamma desempeñan un papel importante en la teoría de colas y en problemas de confiabilidad. Los tiempos entre llegadas en instalaciones de servicio y los tiempos de operación antes de que partes componentes y sistemas eléctricos empiecen a fallar a menudo se representan bien mediante la distribución exponencial. La relación entre la distribución gamma y la exponencial permite que la gamma se utilice en problemas similares. En la siguiente sección se presentarán más detalles y ejemplos.

La distribución gamma deriva su nombre de la bien conocida **función gamma**, que se estudia en muchas áreas de las matemáticas. Antes de estudiar la distribución gamma repasaremos esta función y algunas de sus propiedades importantes.

Definición 6.2: La **función gamma** se define como

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \quad \text{para} \quad \alpha > 0.$$

Las siguientes son algunas propiedades sencillas de la función gamma.

a) $\Gamma(n) = (n-1)(n-2)\cdots(1)\,\Gamma(1)$ para una integral positiva n.

Para ver la demostración, al integrar por partes con $u = x^{\alpha-1}$ y $dv = e^{-x}\, dx$, obtenemos

$$\Gamma(\alpha) = -e^{-x}\, x^{\alpha-1}\Big|_0^\infty + \int_0^\infty e^{-x}(\alpha-1)x^{\alpha-2}\, dx = (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}\, dx,$$

para $\alpha > 1$, que produce la fórmula recursiva

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1).$$

El resultado proviene de la aplicación repetida de la fórmula recursiva. Si utilizamos este resultado, podemos demostrar con facilidad las siguientes dos propiedades.

b) $\Gamma(n) = (n-1)!$ para una integral positiva n.

c) $\Gamma(1) = 1$.

Asimismo, tenemos la siguiente propiedad de $\Gamma(\alpha)$, que el lector deberá verificar (véase el ejercicio 6.39 de la página 206).

d) $\Gamma(1/2) = \sqrt{\pi}$.

A continuación se define la **distribución gamma**.

Distribución gamma

La variable aleatoria continua X tiene una **distribución gamma**, con parámetros α y β, si su función de densidad está dada por

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$.

En la figura 6.28 se muestran gráficas de varias distribuciones gamma para ciertos valores específicos de los parámetros α y β. La distribución gamma especial para la que $\alpha = 1$ se llama **distribución exponencial**.

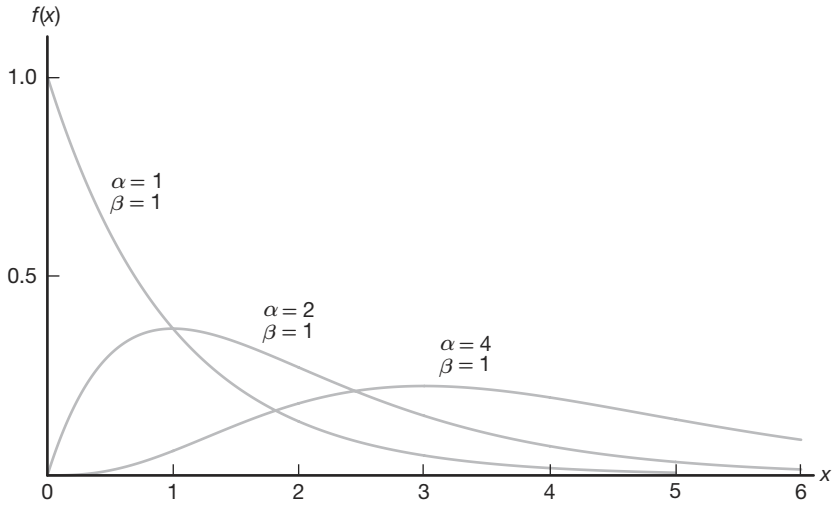


Figura 6.28: Distribuciones gamma.

Distribución exponencial

La variable aleatoria continua X tiene una **distribución exponencial**, con parámetro β, si su función de densidad es dada por

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\beta > 0$.

El siguiente teorema y corolario proporcionan la media y la varianza de la distribución gamma y la exponencial.

Teorema 6.4: La media y la varianza de la distribución gamma son

$$\mu = \alpha\beta \text{ y } \sigma^2 = \alpha\beta^2.$$

La demostración de este teorema se encuentra en el apéndice A.26.

Corolario 6.1: La media y la varianza de la distribución exponencial son

$$\mu = \beta \text{ y } \sigma^2 = \beta^2.$$

Relación con el proceso de Poisson

Continuaremos con las aplicaciones de la distribución exponencial y después regresaremos a la distribución gamma. Las aplicaciones más importantes de la distribución exponencial son situaciones donde se aplica el proceso de Poisson (véase la sección 5.5). El lector debería recordar que el proceso de Poisson permite utilizar la distribución discreta llamada distribución de Poisson. Recuerde que la distribución de Poisson se utiliza para calcular la probabilidad de números específicos de "eventos" durante un *periodo o espacio* particulares. En muchas aplicaciones la variable aleatoria es el tiempo o la cantidad de espacio. Por ejemplo, un ingeniero industrial se podría interesar en un modelo de tiempo T entre las llegadas en una intersección congestionada durante las horas de mayor afluencia en una ciudad grande. Una llegada representa el evento de Poisson.

La relación entre la distribución exponencial (a menudo denominada exponencial negativa) y el proceso de Poisson es muy simple. En el capítulo 5 la distribución de Poisson se desarrolló como una distribución de un solo parámetro con parámetro λ, donde λ se interpreta como el número medio de eventos por *unidad de "tiempo"*. Considere ahora la *variable aleatoria* descrita por el tiempo que se requiere para que ocurra el primer evento. Si utilizamos la distribución de Poisson, vemos que la probabilidad de que no ocurra algún evento, en el periodo hasta el tiempo t, es dada por

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Ahora podemos utilizar lo anterior y hacer que X sea el tiempo para el primer evento de Poisson. La probabilidad de que la duración del tiempo hasta el primer evento exceda x es la misma que la probabilidad de que no ocurra algún evento de Poisson en x. Esto último, por supuesto, es dado por $e^{-\lambda x}$. Como resultado,

$$P(X > x) = e^{-\lambda x}.$$

Así, la función de distribución acumulativa para X es dada por

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x}.$$

Ahora, para poder reconocer la presencia de la distribución exponencial, podemos diferenciar la función de distribución acumulativa anterior con el fin de obtener la función de densidad

$$f(x) = \lambda e^{-\lambda x},$$

que es la función de densidad de la distribución exponencial con $\lambda = 1/\beta$.

Aplicaciones de la distribución exponencial y la distribución gamma

En la explicación anterior establecimos las bases para la aplicación de la distribución exponencial en el "tiempo de llegada" o tiempo para problemas con eventos de Poisson. Aquí ilustraremos algunas aplicaciones de modelado y después procederemos a analizar el papel que la distribución gamma desempeña en ellas. Observe que la media de la distribución exponencial es el parámetro β, el recíproco del parámetro en la distribución de Poisson. El lector debería recordar que con frecuencia se dice que la distribución de Poisson no tiene memoria, lo cual implica que las ocurrencias en periodos sucesivos son independientes. El importante parámetro β es el tiempo promedio entre eventos. En la teoría de confiabilidad, donde la falla de equipo con frecuencia se ajusta a este proceso de Poisson, β se denomina **tiempo medio entre fallas**. Muchas descomposturas de equipo siguen el proceso de Poisson y por ello se aplica la distribución exponencial. Otras aplicaciones incluyen tiempos de supervivencia en experimentos biomédicos y tiempo de respuesta de computadoras.

En el siguiente ejemplo mostramos una aplicación simple de la distribución exponencial a un problema de confiabilidad. La distribución binomial también desempeña un papel en la solución.

Ejemplo 6.17: Suponga que un sistema contiene cierto tipo de componente cuyo tiempo de operación antes de fallar, en años, está dado por T. La variable aleatoria T se modela bien mediante la distribución exponencial con tiempo medio de operación antes de fallar $\beta = 5$. Si se instalan 5 de estos componentes en diferentes sistemas, ¿cuál es la probabilidad de que al final de 8 años al menos dos aún funcionen?

***Solución*:** La probabilidad de que un componente determinado siga funcionando después de 8 años es dada por

$$P(T > 8) = \frac{1}{5}\int_8^\infty e^{-t/5}\, dt = e^{-8/5} \approx 0.2.$$

Representemos con X el número de componentes que todavía funcionan después de 8 años. Entonces, utilizando la distribución binomial tenemos

$$P(X \geq 2) = \sum_{x=2}^{5} b(x; 5, 0.2) = 1 - \sum_{x=0}^{1} b(x; 5, 0.2) = 1 - 0.7373 = 0.2627.$$

En el capítulo 3 se incluyen ejercicios y ejemplos en los que el lector ya se enfrentó a la distribución exponencial. Otros que implican problemas de tiempo de espera y de confiabilidad se pueden encontrar en el ejemplo 6.24 y en los ejercicios y ejercicios de repaso al final de este capítulo.

La propiedad de falta de memoria y su efecto en la distribución exponencial

En los tipos de aplicación de la distribución exponencial en los problemas de confiabilidad y de tiempo de vida de una máquina o de un componente influye la **propiedad de**

falta de memoria de la distribución exponencial. Por ejemplo, en el caso de, digamos, un componente electrónico, en el que la distribución del tiempo de vida es exponencial, la probabilidad de que el componente dure, por ejemplo, t horas, es decir, $P(X \geq t)$, es igual que la probabilidad condicional

$$P(X \geq t_0 + t \mid X \geq t_0).$$

Entonces, si el componente "alcanza" las t_0 horas, la probabilidad de que dure otras t horas es igual que la probabilidad de que dure t horas. No hay "castigo" a través del desgaste como resultado de durar las primeras t_0 horas. Por lo tanto, cuando la propiedad de falta de memoria es justificada es más adecuada la distribución exponencial. Pero si la falla del componente es resultado del desgaste lento o gradual (como en el caso del desgaste mecánico), entonces la distribución exponencial no es aplicable y serían más adecuadas la distribución gamma o la de Weibull (sección 6.10).

La importancia de la distribución gamma radica en el hecho de que define una familia en la cual otras distribuciones son casos especiales. Pero la propia distribución gamma tiene aplicaciones importantes en tiempo de espera y teoría de confiabilidad. Mientras que la distribución exponencial describe el tiempo que transcurre hasta la ocurrencia de un evento de Poisson (o el tiempo entre eventos de Poisson), el tiempo (o espacio) que transcurre hasta que *ocurre un número específico de eventos de Poisson* es una variable aleatoria, cuya función de densidad es descrita por la distribución gamma. Este número específico de eventos es el parámetro α en la función de densidad gamma. De esta manera se facilita comprender que cuando $\alpha = 1$, ocurre el caso especial de la distribución exponencial. La densidad gamma se puede desarrollar a partir de su relación con el proceso de Poisson de la misma manera en que lo hicimos con la densidad exponencial. Los detalles se dejan al lector. El siguiente es un ejemplo numérico de cómo se utiliza la distribución gamma en una aplicación de tiempo de espera.

Ejemplo 6.18: Suponga que las llamadas telefónicas que llegan a un conmutador particular siguen un proceso de Poisson con un promedio de 5 llamadas entrantes por minuto. ¿Cuál es la probabilidad de que transcurra hasta un minuto en el momento en que han entrado 2 llamadas al conmutador?

***Solución*:** Se aplica el proceso de Poisson, con un lapso de tiempo hasta que ocurren 2 eventos de Poisson que sigue una distribución gamma con $\beta = 1/5$ y $\alpha = 2$. Denote con X el tiempo en minutos que transcurre antes de que lleguen 2 llamadas. La probabilidad que se requiere está dada por

$$P(X \leq 1) = \int_0^1 \frac{1}{\beta^2} x e^{-x/\beta}\, dx = 25 \int_0^1 x e^{-5x}\, dx = 1 - e^{-5}(1 + 5) = 0.96.$$

Mientras el origen de la distribución gamma trata con el tiempo (o espacio) hasta la ocurrencia de α eventos de Poisson, hay muchos ejemplos donde una distribución gamma funciona muy bien aunque no exista una estructura de Poisson clara. Esto es particularmente cierto para problemas de **tiempo de supervivencia** en aplicaciones de ingeniería y biomédicas.

Ejemplo 6.19: En un estudio biomédico con ratas se utiliza una investigación de respuesta a la dosis para determinar el efecto de la dosis de un tóxico en su tiempo de supervivencia. El tóxico es producido por el combustible que utilizan los aviones y, en consecuencia, descargan con frecuencia a la atmósfera. Para cierta dosis del tóxico, el estudio determina que el tiempo de supervivencia de las ratas, en semanas, tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$. ¿Cuál es la probabilidad de que una rata no sobreviva más de 60 semanas?

***Solución*:** Sea la variable aleatoria X el tiempo de supervivencia (tiempo hasta la muerte). La probabilidad que se requiere es

$$P(X \leq 60) = \frac{1}{\beta^5} \int_0^{60} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(5)}\, dx.$$

La integral anterior se puede resolver mediante la **función gamma incompleta**, que se convierte en la función de distribución acumulativa para la distribución gamma. Esta función se escribe como

$$F(x;\alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy.$$

Si permitimos que $y = x/\beta$, de modo que $x = \beta y$, tenemos

$$P(X \leq 60) = \int_0^6 \frac{y^4 e^{-y}}{\Gamma(5)}\, dy,$$

que se denota como $F(6; 5)$ en la tabla de la función gamma incompleta del apéndice A.23. Observe que esto permite un cálculo rápido de las probabilidades para la distribución gamma. De hecho, para este problema la probabilidad de que la rata no sobreviva más de 60 días es dada por

$$P(X \leq 60) = F(6;5) = 0.715.$$

Ejemplo 6.20: A partir de datos previos se sabe que la longitud de tiempo, en meses, entre las quejas de los clientes sobre cierto producto es una distribución gamma con $\alpha = 2$ y $\beta = 4$. Se realizaron cambios para hacer más estrictos los requerimientos del control de calidad después de los cuales pasaron 20 meses antes de la primera queja. ¿Parecería que los cambios realizados en el control de calidad resultaron eficaces?

***Solución*:** Sea X el tiempo para que se presente la primera queja, el cual, en las condiciones anteriores a los cambios, seguía una distribución gamma con $\alpha = 2$ y $\beta = 4$. La pregunta se centra alrededor de qué tan raro es $X \geq 20$ dado que α y β permanecen con los valores 2 y 4, repectivamente. En otras palabras, en las condiciones anteriores ¿es razonable un "tiempo para la queja" tan grande como 20 meses? Por consiguiente, si seguimos la solución del ejemplo 6.19,

$$P(X \geq 20) = 1 - \frac{1}{\beta^\alpha} \int_0^{20} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} dx.$$

De nuevo, usando $y = x/\beta$ tenemos

$$P(X \geq 20) = 1 - \int_0^5 \frac{y e^{-y}}{\Gamma(2)}\, dy = 1 - F(5;2) = 1 - 0.96 = 0.04,$$

donde $F(5; 2) = 0.96$ se obtiene de la tabla A.23.

Como resultado, podríamos concluir que las condiciones de la distribución gamma con $\alpha = 2$ y $\beta = 4$ no son sustentadas por los datos de que un tiempo observado para la queja sea tan extenso como 20 meses. Entonces, es razonable concluir que el trabajo de control de calidad resultó eficaz.

Ejemplo 6.21: Considere el ejercicio 3.31 de la página 94. Con base en abundantes pruebas se determinó que el tiempo Y en años antes de que se requiera una reparación mayor para cierta lavadora se caracteriza por la función de densidad

$$f(y) = \begin{cases} \frac{1}{4} e^{-y/4}, & y \geq 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Observe que Y es una variable aleatoria exponencial con $\mu = 4$ años. Se considera que la lavadora es una ganga si no hay probabilidades de que requiera una reparación mayor antes de cumplir 6 años de haber sido comprada. ¿Cuál es la probabilidad de $P(Y > 6)$? ¿Cuál es la probabilidad de que la lavadora requiera una reparación mayor durante el primer año?

***Solución*:** Considere la función de distribución acumulativa $F(y)$ para la distribución exponencial,

$$F(y) = \frac{1}{\beta}\int_0^y e^{-t/\beta}\,dt = 1 - e^{-y/\beta}.$$

De manera que

$$P(Y > 6) = 1 - F(6) = e^{-3/2} = 0.2231.$$

Por lo tanto, la probabilidad de que la lavadora requiera una reparación mayor después de seis años es de 0.223. Desde luego, la probabilidad de que requiera reparación antes del sexto año es de 0.777. Así, se podría concluir que la lavadora no es realmente una ganga. La probabilidad de que se requiera una reparación mayor durante el primer año es

$$P(Y < 1) = 1 - e^{-1/4} = 1 - 0.779 = 0.221.$$

6.7 Distribución chi cuadrada

Otro caso especial muy importante de la distribución gamma se obtiene al permitir que $\alpha = v/2$ y $\beta = 2$, donde v es un entero positivo. Este resultado se conoce como **distribución chi cuadrada**. La distribución tiene un solo parámetro, v, denominado **grados de libertad**.

Distribución chi cuadrada

La variable aleatoria continua X tiene una **distribución chi cuadrada**, con v **grados de libertad**, si su función de densidad es dada por

$$f(x; v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde v es un entero positivo.

La distribución chi cuadrada desempeña un papel fundamental en la inferencia estadística. Tiene una aplicación considerable tanto en la metodología como en la teoría. Aunque no estudiaremos con detalle sus aplicaciones en este capítulo, es importante tener en cuenta que los capítulos 8, 9 y 16 contienen aplicaciones importantes. La distribución chi cuadrada es un componente importante de la prueba estadística de hipótesis y de la estimación estadística.

Los temas en los que se trata con distribuciones de muestreo, análisis de varianza y estadística no paramétrica implican el uso extenso de la distribución chi cuadrada.

Teorema 6.5: La media y la varianza de la distribución chi cuadrada son

$$\mu = v \text{ y } \sigma^2 = 2v.$$

6.8 Distribución beta

Una extensión de la distribución uniforme es la distribución beta. Primero definiremos una **función beta**.

Definición 6.3: Una **función beta** es definida por

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},\ \text{para } \alpha,\beta>0,$$

donde $\Gamma(\alpha)$ es la función gamma.

Distribución beta

La variable aleatoria continua X tiene una **distribución beta** con los parámetros $\alpha>0$ y $\beta>0$, si su función de densidad es dada por

$$f(x)=\begin{cases}\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1,\\ 0, & \text{en otro caso.}\end{cases}$$

Observe que la distribución uniforme sobre (0, 1) es una distribución beta con los parámetros $\alpha=1$ y $\beta=1$.

Teorema 6.6: La media y la varianza de una distribución beta en la que los parámetros α y β son

$$\mu=\frac{\alpha}{\alpha+\beta}\ \text{y}\ \sigma^2=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectivamente.

Para la distribución uniforme sobre (0, 1), la media y la varianza son

$$\mu=\frac{1}{1+1}=\frac{1}{2}\ \text{y}\ \sigma^2=\frac{(1)(1)}{(1+1)^2(1+1+1)}=\frac{1}{12},$$

respectivamente.

6.9 Distribución logarítmica normal

La distribución logarítmica normal se utiliza en una amplia variedad de aplicaciones. La distribución se aplica en casos donde una transformación logarítmica natural tiene como resultado una distribución normal.

Distribución logarítmica normal

La variable aleatoria continua X tiene una **distribución logarítmica normal** si la variable aleatoria $Y=\ln(X)$ tiene una distribución normal con media μ y desviación estándar σ. La función de densidad de X que resulta es

$$f(x;\mu,\sigma)=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x\geq 0,\\ 0, & x<0.\end{cases}$$

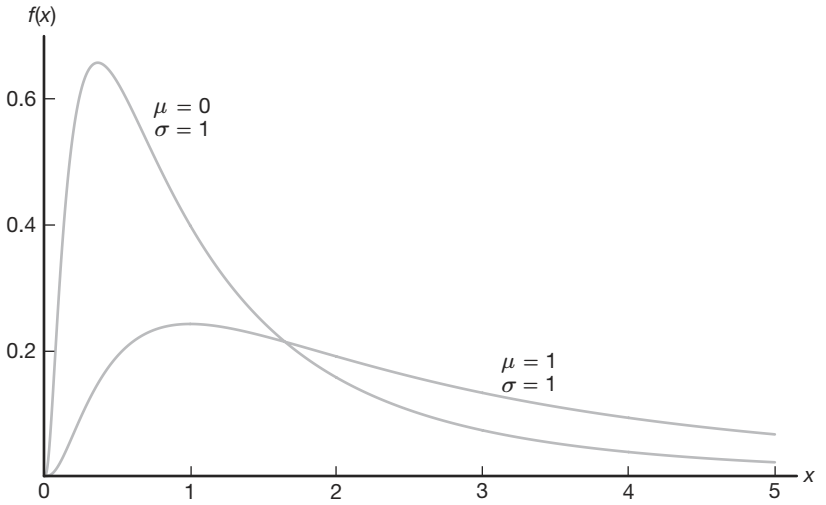


Figura 6.29: Distribuciones logarítmicas normales.

Las gráficas de las distribuciones logarítmicas normales se ilustran en la figura 6.29.

Teorema 6.7: La media y la varianza de la distribución logarítmica normal son

$$\mu = e^{\mu+\sigma^2/2} \text{ y } \sigma^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

La función de distribución acumulativa es muy simple debido a su relación con la distribución normal. El uso de la función de distribución se ilustra con el siguiente ejemplo.

Ejemplo 6.22: Se sabe que históricamente la concentración de contaminantes producidos por plantas químicas exhiben un comportamiento que se parece a una distribución logarítmica normal. Esto es importante cuando se consideran cuestiones relacionadas con el cumplimiento de las regulaciones gubernamentales. Suponga que la concentración de cierto contaminante, en partes por millón, tiene una distribución logarítmica normal con los parámetros $\mu = 3.2$ y $\sigma = 1$. ¿Cuál es la probabilidad de que la concentración exceda 8 partes por millón?

***Solución*:** Sea la variable aleatoria X la concentración de contaminantes. Entonces

$$P(X > 8) = 1 - P(X \leq 8).$$

Como $\ln(X)$ tiene una distribución normal con media $\mu = 3.2$ y desviación estándar $\sigma = 1$,

$$P(X \leq 8) = \Phi\left[\frac{\ln(8) - 3.2}{1}\right] = \Phi(-1.12) = 0.1314.$$

Aquí, utilizamos el símbolo Φ para denotar la función de distribución acumulativa de la distribución normal estándar. Como resultado, la probabilidad de que la concentración del contaminante exceda 8 partes por millón es 0.1314.

Ejemplo 6.23: La vida, en miles de millas, de un cierto tipo de control electrónico para locomotoras tiene una distribución aproximadamente logarítmica normal con $\mu = 5.149$ y $\sigma = 0.737$. Calcule el quinto percentil de la vida de un control electrónico como éste.

***Solución*:** A partir de la tabla A.3 sabemos que $P(Z < -1.645) = 0.05$. Denote como X la vida del control electrónico. Puesto que $\ln(X)$ tiene una distribución normal con media $\mu = 5.149$ y $\sigma = 0.737$, el quinto percentil de X se calcula como

$$\ln(x) = 5.149 + (0.737)(-1.645) = 3.937.$$

Por lo tanto, $x = 51.265$. Esto significa que sólo 5% de los controles tendrán un tiempo de vida menor que 51,265 millas. ∎

6.10 Distribución de Weibull (opcional)

La tecnología actual permite que los ingenieros diseñen muchos sistemas complicados cuya operación y seguridad dependen de la confiabilidad de los diversos componentes que conforman los sistemas. Por ejemplo, un fusible se puede quemar, una columna de acero se puede torcer o un dispositivo sensor de calor puede fallar. Componentes idénticos, sujetos a idénticas condiciones ambientales, fallarán en momentos diferentes e impredecibles. Ya examinamos el papel que desempeñan las distribuciones gamma y exponencial en estos tipos de problemas. Otra distribución que se ha utilizado ampliamente en años recientes para tratar con tales problemas es la **distribución de Weibull**, introducida por el físico sueco Waloddi Weibull en 1939.

Distribución de Weibull

La variable aleatoria continua X tiene una **distribución de Weibull**, con parámetros α y β, si su función de densidad es dada por

$$f(x;\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$.

En la figura 6.30 se ilustran las gráficas de la distribución de Weibull para $\alpha = 1$ y diversos valores del parámetro β. Vemos que las curvas cambian de manera considerable para diferentes valores del parámetro β. Si permitimos que $\beta = 1$, la distribución de Weibull se reduce a la distribución exponencial. Para valores de $\beta > 1$ las curvas adoptan ligeramente la forma de campana y se asemejan a las curvas normales, pero muestran algo de asimetría.

La media y la varianza de la distribución de Weibull se establecen en el siguiente teorema. Se solicita al lector que haga la demostración en el ejercicio 6.52 de la página 206.

Teorema 6.8: La media y la varianza de la distribución de Weibull son

$$\mu = \alpha^{-1/\beta}\Gamma\left(1 + \frac{1}{\beta}\right) \text{ y } \sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right\}.$$

Al igual que la distribución gamma y la exponencial, la distribución de Weibull se aplica a problemas de confiabilidad y de prueba de vida como los de **tiempo de operación**

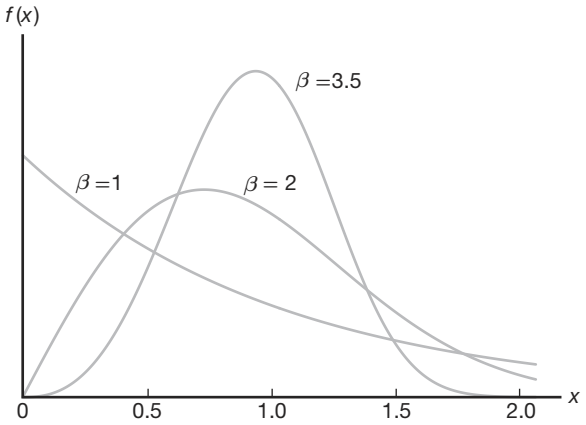


Figura 6.30: Distribuciones de Weibull ($\alpha = 1$).

antes de la falla o la **duración de la vida** de un componente, que se miden desde algún tiempo específico hasta que falla. Representemos este tiempo de operación antes de la falla mediante la variable aleatoria continua T, con función de densidad de probabilidad $f(t)$, donde $f(t)$ es la distribución de Weibull. Ésta tiene la flexibilidad inherente de no requerir la propiedad de falta de memoria de la distribución exponencial. La función de distribución acumulativa (fda) para la distribución de Weibull se puede escribir en forma cerrada y realmente es muy útil para calcular probabilidades.

Fda para la distribución de Weibull

La **función de distribución acumulativa para la distribución de Weibull** es dada por

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \quad \text{para} \quad x \geq 0,$$

para $\alpha > 0$ y $\beta > 0$.

Ejemplo 6.24: El tiempo de vida X, en horas, de un artículo en el taller mecánico tiene una distribución de Weibull con $\alpha = 0.01$ y $\beta = 2$. ¿Cuál es la probabilidad de que falle antes de 8 horas de uso?

***Solución*:** $P(X < 8) = F(8) = 1 - e^{-(0.01)8^2} = 1 - 0.527 = 0.473.$

La tasa de fallas para la distribución de Weibull

Cuando se aplica la distribución de Weibull, con frecuencia es útil determinar la **tasa de fallas** (algunas veces denominada tasa de riesgo) para tener conocimiento del desgaste o deterioro del componente. Comencemos por definir la confiabilidad de un componente o producto como la *probabilidad de que funcione adecuadamente por al menos un tiempo específico en condiciones experimentales específicas*. Por lo tanto, si $R(t)$ se define como la confiabilidad del componente dado en el tiempo t, escribimos

$$R(t) = P(T > t) = \int_{t}^{\infty} f(t)\, dt = 1 - F(t),$$

donde $F(t)$ es la función de distribución acumulativa de T. La probabilidad condicional de que un componente fallará en el intervalo de $T = t$ a $T = t + \Delta t$, dado que sobrevive hasta el tiempo t, es

$$\frac{F(t + \Delta t) - F(t)}{R(t)}.$$

Al dividir esta proporción entre Δt y tomar el límite como $\Delta t \to 0$, obtenemos la **tasa de fallas**, denotada por $Z(t)$. De aquí,

$$Z(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)},$$

que expresa la tasa de fallas en términos de la distribución del tiempo de operación antes de la falla.

Como $Z(t) = f(t)/[1 - F(t)]$, entonces la tasa de falla es dada como sigue:

Tasa de fallas para la distribución de Weibull

La **tasa de fallas** en el tiempo t para la distribución de Weibull es dada por

$$Z(t) = \alpha\beta t^{\beta - 1}, \qquad t > 0.$$

Interpretación de la tasa de fallas

La cantidad $Z(t)$ es bien llamada tasa de fallas porque realmente cuantifica la tasa de cambio con el tiempo de la probabilidad condicional de que el componente dure una Δt adicional *dado que ha durado el tiempo t*. La tasa de disminución (o crecimiento) con el tiempo también es importante. Los siguientes puntos son fundamentales.

a) Si $\beta = 1$, la tasa de fallas $= \alpha$, es decir, una constante. Esto, como se indicó anteriormente, es el caso especial de la distribución exponencial en que predomina la falta de memoria.

b) Si $\beta > 1$, $Z(t)$ es una función creciente del tiempo t que indica que el componente se desgasta con el tiempo.

c) Si $\beta < 1$, $Z(t)$ es una función decreciente del tiempo t y, por lo tanto, el componente se fortalece o endurece con el paso del tiempo.

Por ejemplo, el artículo en el taller mecánico del ejemplo 6.24 tiene $\beta = 2$ y, por consiguiente, se desgasta con el tiempo. De hecho, la función de la tasa de fallas es dada por $Z(t) = .02t$. Por otro lado, suponga un parámetro donde $\beta = 3/4$ y $\alpha = 2$. En ese caso, $Z(t) = 1.5/t^{1/4}$ y, por lo tanto, el componente se hace más fuerte con el tiempo.

Ejercicios

6.39 Utilice la función gamma con $y = \sqrt{2x}$ para demostrar que $\Gamma(1/2) = \sqrt{\pi}$.

6.40 En cierta ciudad, el consumo diario de agua (en millones de litros) sigue aproximadamente una distribución gamma con $\alpha = 2$ y $\beta = 3$. Si la capacidad diaria de dicha ciudad es de 9 millones de litros de agua, ¿cuál es la probabilidad de que en cualquier día dado el suministro de agua sea inadecuado?

6.41 Si una variable aleatoria X tiene una distribución gamma con $\alpha = 2$ y $\beta = 1$, calcule $P(1.8 < X < 2.4)$.

6.42 Suponga que el tiempo, en horas, necesario para reparar una bomba de calor es una variable aleatoria X que tiene una distribución gamma con los parámetros $\alpha = 2$ y $\beta = 1/2$. ¿Cuál es la probabilidad de que la siguiente llamada de servicio requiera

a) a lo sumo una hora para reparar la bomba de calor?
b) al menos dos horas para reparar la bomba de calor?

6.43 *a*) Calcule la media y la varianza del consumo diario de agua del ejercicio 6.40.

b) De acuerdo con el teorema de Chebyshev, ¿hay por lo menos $3/4$ de probabilidad de que el consumo de agua en cualquier día determinado caiga dentro de cuál intervalo?

6.44 En cierta ciudad el consumo diario de energía eléctrica, en millones de kilowatts-hora, es una variable aleatoria X que tiene una distribución gamma con media $\mu = 6$ y varianza $\sigma^2 = 12$.

a) Calcule los valores de α y β.
b) Calcule la probabilidad de que en cualquier día dado el consumo diario de energía exceda los 12 millones de kilowatts-hora.

6.45 El tiempo necesario para que un individuo sea atendido en una cafetería es una variable aleatoria que tiene una distribución exponencial con una media de 4 minutos. ¿Cuál es la probabilidad de que una persona sea atendida en menos de 3 minutos en al menos 4 de los siguientes 6 días?

6.46 La vida, en años, de cierto interruptor eléctrico tiene una distribución exponencial con una vida promedio de $\beta = 2$. Si 100 de estos interruptores se instalan en diferentes sistemas, ¿cuál es la probabilidad de que, a lo sumo, fallen 30 durante el primer año?

6.47 Suponga que la vida de servicio de la batería de un auxiliar auditivo, en años, es una variable aleatoria que tiene una distribución de Weibull con $\alpha = 1/2$ y $\beta = 2$.

a) ¿Cuánto tiempo se puede esperar que dure tal batería?
b) ¿Cuál es la probabilidad de que tal batería esté funcionando después de 2 años?

6.48 Derive la media y la varianza de la distribución beta.

6.49 Suponga que la variable aleatoria X tiene una distribución beta con $\alpha = 1$ y $\beta = 3$.

a) Determine la media y la mediana de X.
b) Determine la varianza de X.
c) Calcule la probabilidad de que $X > 1/3$.

6.50 Si la proporción de una marca de televisores que requiere servicio durante el primer año de operación es una variable aleatoria que tiene una distribución beta con $\alpha = 3$ y $\beta = 2$, ¿cuál es la probabilidad de que al menos 80% de los nuevos modelos de esta marca que se vendieron este año requieran servicio durante su primer año de operación?

6.51 Las vidas de ciertos sellos para automóvil tienen la distribución de Weibull con tasa de fallas $Z(t) = 1/\sqrt{t}$. Calcule la probabilidad de que tal sello aún esté intacto después de 4 años.

6.52 Derive la media y la varianza de la distribución de Weibull.

6.53 En una investigación biomédica se determinó que el tiempo de supervivencia, en semanas, de un animal cuando se le somete a cierta exposición de radiación gamma tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$.

a) ¿Cuál es el tiempo medio de supervivencia de un animal seleccionado al azar del tipo que se utilizó en el experimento?
b) ¿Cuál es la desviación estándar del tiempo de supervivencia?
c) ¿Cuál es la probabilidad de que un animal sobreviva más de 30 semanas?

6.54 Se sabe que la vida, en semanas, de cierto tipo de transistor tiene una distribución gamma con una media de 10 semanas y una desviación estándar de $\sqrt{50}$ semanas.

a) ¿Cuál es la probabilidad de que un transistor de este tipo dure a lo sumo 50 semanas?
b) ¿Cuál es la probabilidad de que un transistor de este tipo no sobreviva las primeras 10 semanas?

6.55 El tiempo de respuesta de una computadora es una aplicación importante de las distribuciones gamma y exponencial. Suponga que un estudio de cierto sistema de cómputo revela que el tiempo de respuesta, en segundos, tiene una distribución exponencial con una media de 3 segundos.

a) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 5 segundos?
b) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 10 segundos?

6.56 Los datos de frecuencia a menudo tienen una distribución logarítmica normal. Se estudia el uso promedio de potencia (dB por hora) para una empresa específica y se sabe que tiene una distribución logarítmica normal con parámetros $\mu = 4$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la empresa utilice más de 270 dB durante cualquier hora particular?

6.57 Para el ejercicio 6.56, ¿cuál es el uso de la potencia media (dB promedio por hora)? ¿Cuál es la varianza?

6.58 El número de automóviles que llegan a cierta intersección por minuto tiene una distribución de Poisson con una media de 5. Existe interés por el tiempo que transcurre antes de que 10 automóviles aparezcan en la intersección.

a) ¿Cuál es la probabilidad de que más de 10 automóviles aparezcan en la intersección durante cualquier minuto determinado?
b) ¿Cuál es la probabilidad de que transcurran más de 2 minutos antes de que lleguen 10 autos?

6.59 Considere la información del ejercicio 6.58.
a) ¿Cuál es la probabilidad de que transcurra más de 1 minuto entre llegadas?
b) ¿Cuál es el número medio de minutos que transcurre entre las llegadas?

6.60 Demuestre que la función de la tasa de fallas es dada por

$$Z(t) = \alpha\beta t^{\beta-1}, \quad t > 0,$$

si y sólo si la distribución del tiempo que transcurre antes de la falla es la distribución de Weibull

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}, \quad t > 0.$$

Ejercicios de repaso

6.61 Según un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, aproximadamente 49% de los consumidores de Valium en el estado de Massachusetts son empleados de oficina. ¿Cuál es la probabilidad de que entre 482 y 510 de los siguientes 1000 consumidores de Valium seleccionados al azar de dicho estado sean empleados de oficina?

6.62 La distribución exponencial se aplica con frecuencia a los tiempos de espera entre éxitos en un proceso de Poisson. Si el número de llamadas que se reciben por hora en un servicio de respuesta telefónica es una variable aleatoria de Poisson con el parámetro $\lambda = 6$, sabemos que el tiempo, en horas, entre llamadas sucesivas tiene una distribución exponencial con el parámetro $\beta = 1/6$. ¿Cuál es la probabilidad de esperar más de 15 minutos entre cualesquiera 2 llamadas sucesivas?

6.63 Cuando α es un entero positivo n, la distribución gamma también se conoce como **distribución de Erlang**. Al establecer que $\alpha = n$ en la distribución gamma de la página 195, la distribución de Erlang es

$$f(x) = \begin{cases} \frac{x^{n-1} e^{-x/\beta}}{\beta^n (n-1)!}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Se puede demostrar que si los tiempos entre eventos sucesivos son independientes, y cada uno tiene una distribución exponencial con el parámetro β, entonces el tiempo de espera total X transcurrido hasta que ocurran n eventos tiene la distribución de Erlang. Con referencia al ejercicio de repaso 6.62, ¿cuál es la probabilidad de que las siguientes 3 llamadas se reciban dentro de los siguientes 30 minutos?

6.64 Un fabricante de cierto tipo de máquina grande desea comprar remaches de uno de dos fabricantes. Es importante que la resistencia a la rotura de cada remache exceda 10,000 psi. Dos fabricantes (A y B) ofrecen este tipo de remache y ambos tienen remaches cuya resistencia a la rotura está distribuida de forma normal. Las resistencias promedio a la rotura para los fabricantes A y B son 14,000 psi y 13,000 psi, respectivamente. Las desviaciones estándar son 2000 psi y 1000 psi, respectivamente. ¿Cuál fabricante producirá, en promedio, el menor número de remaches defectuosos?

6.65 De acuerdo con un censo reciente, casi 65% de los hogares en Estados Unidos se componen de una o dos personas. Si se supone que este porcentaje sigue siendo válido en la actualidad, ¿cuál es la probabilidad de que entre 590 y 625 de los siguientes 1000 hogares seleccionados al azar en Estados Unidos consten de una o dos personas?

6.66 Cierto tipo de dispositivo tiene una tasa de fallas anunciada de 0.01 por hora. La tasa de fallas es constante y se aplica la distribución exponencial.
a) ¿Cuál es el tiempo promedio que transcurre antes de la falla?
b) ¿Cuál es la probabilidad de que pasen 200 horas antes de que se observe una falla?

6.67 En una planta de procesamiento químico es importante que el rendimiento de cierto tipo de producto de un lote se mantenga por arriba de 80%. Si permanece por debajo de 80% durante un tiempo prolongado, la empresa pierde dinero. Los lotes producidos ocasionalmente con defectos son de poco interés, pero si varios lotes por día resultan defectuosos, la planta se detiene y se realizan ajustes. Se sabe que el rendimiento se distribuye normalmente con una desviación estándar de 4%.

a) ¿Cuál es la probabilidad de una "falsa alarma" (rendimiento por debajo de 80%) cuando el rendimiento promedio es en realidad de 85%?

b) ¿Cuál es la probabilidad de que un lote tenga un rendimiento que exceda el 80% cuando en realidad el rendimiento promedio es de 79%?

6.68 Para un componente eléctrico que tiene una tasa de fallas de una vez cada 5 horas es importante considerar el tiempo que transcurre para que fallen 2 componentes.

a) Suponiendo que se aplica la distribución gamma, ¿cuál es el tiempo promedio que transcurre para que fallen 2 componentes?

b) ¿Cuál es la probabilidad de que transcurran 12 horas antes de que fallen 2 componentes?

6.69 Se establece que la elongación de una barra de acero bajo una carga particular se distribuye normalmente con una media de 0.05 pulgadas y $\sigma = 0.01$ pulgadas. Calcule la probabilidad de que el alargamiento esté

a) por arriba de 0.1 pulgadas;

b) por abajo de 0.04 pulgadas;

c) entre 0.025 y 0.065 pulgadas.

6.70 Se sabe que un satélite controlado tiene un error (distancia del objetivo) que se distribuye normalmente con una media 0 y una desviación estándar de 4 pies. El fabricante del satélite define un éxito como un disparo en el cual el satélite llega a 10 pies del objetivo. Calcule la probabilidad de que el satélite falle.

6.71 Un técnico planea probar cierto tipo de resina desarrollada en el laboratorio para determinar la naturaleza del tiempo que transcurre antes de que se logre el pegado. Se sabe que el tiempo promedio para el pegado es de 3 horas y que la desviación estándar es de 0.5 horas. Un producto se considerará indeseable si el tiempo de pegado es menor de una hora o mayor de 4 horas. Comente sobre la utilidad de la resina. ¿Con qué frecuencia su desempeño se considera indeseable? Suponga que el tiempo para la unión se distribuye normalmente.

6.72 Considere la información del ejercicio de repaso 6.66. ¿Cuál es la probabilidad de que transcurran menos de 200 horas antes de que ocurran 2 fallas?

6.73 Para el ejercicio de repaso 6.72, ¿cuál es la media y la varianza del tiempo que transcurre antes de que ocurran 2 fallas?

6.74 Se sabe que la tasa promedio de uso de agua (en miles de galones por hora) en cierta comunidad implica la distribución logarítmica normal con los parámetros $\mu = 5$ y $\sigma = 2$. Para propósitos de planeación es importante tener información sobre los periodos de alto consumo. ¿Cuál es la probabilidad de que, para cualquier hora determinada, se usen 50,000 galones de agua?

6.75 Para el ejercicio de repaso 6.74, ¿cuál es la media del uso de agua por hora promedio en miles de galones?

6.76 En el ejercicio 6.54 de la página 206 se supone que la vida de un transistor tiene una distribución gamma con una media de 10 semanas y una desviación estándar de $\sqrt{50}$ semanas. Suponga que la distribución gamma es incorrecta y que se trata de una distribución normal.

a) ¿Cuál es la probabilidad de que el transistor dure a lo sumo 50 semanas?

b) ¿Cuál es la probabilidad de que el transistor no sobreviva las primeras 10 semanas?

c) Comente acerca de la diferencia entre los resultados que obtuvo aquí y los que se obtuvieron en el ejercicio 6.54 de la página 206.

6.77 La distribución beta tiene muchas aplicaciones en problemas de confiabilidad, donde la variable aleatoria básica es una proporción, como sucede en el contexto práctico que se ilustra en el ejercicio 6.50 de la página 206. En este apartado considere el ejercicio de repaso 3.73 de la página 108. Las impurezas en el lote del producto de un proceso químico reflejan un problema grave. Se sabe que la proporción de impurezas Y en un lote tiene la siguiente función de densidad

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique que la anterior sea una función de densidad válida.

b) ¿Cuál es la probabilidad de que un lote se considere no aceptable (es decir, $Y > 0.6$)?

c) ¿Cuáles son los parámetros α y β de la distribución beta que se ilustra aquí?

d) La media de la distribución beta es $\frac{\alpha}{\alpha+\beta}$ ¿Cuál es la proporción media de impurezas en el lote?

e) La varianza de una variable aleatoria beta distribuida es

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

¿Cuál es la varianza de Y en este problema?

6.78 Considere ahora el ejercicio de repaso 3.74 de la página 108. La función de densidad del tiempo Z entre las llamadas, en minutos, a una empresa de suministro eléctrico es dada por

$$f(z) = \begin{cases} \frac{1}{10}e^{-z/10}, & 0 < z < \infty, \\ 0, & \text{en otro caso.} \end{cases}$$

a) ¿Cuál es el tiempo medio entre llamadas?
b) ¿Cuál es la varianza en el tiempo entre llamadas?
c) ¿Cuál es la probabilidad de que el tiempo entre llamadas supere la media?

6.79 Considere el ejercicio de repaso 6.78. Dada la suposición de la distribución exponencial, ¿cuál es el número medio de llamadas por hora? ¿Cuál es la varianza en el número de llamadas por hora?

6.80 En un proyecto experimental sobre el factor humano se determinó que el tiempo de reacción de un piloto ante un estímulo visual es distribuido normalmente con una media de $1/2$ segundo y una desviación estándar de $2/5$ de segundo.
a) ¿Cuál es la probabilidad de que una reacción del piloto tome más de 0.3 segundos?
b) ¿Qué tiempo de reacción se excede el 95% de las veces?

6.81 El tiempo que transcurre entre las fallas de una pieza esencial de equipo es importante en la decisión del uso de equipo auxiliar. Un ingeniero cree que el mejor modelo para el tiempo entre las fallas de un generador es la distribución exponencial con una media de 15 días.
a) Si el generador acaba de fallar, ¿cuál es la probabilidad de que falle en los siguientes 21 días?
b) ¿Cuál es la probabilidad de que el generador funcione durante 30 días sin fallar?

6.82 El periodo de vida de una broca en una operación mecánica, en horas, tiene una distribución de Weibull con $\alpha = 2$ y $\beta = 50$. Calcule la probabilidad de que la broca falle antes de 10 horas de uso.

6.83 Calcule la fda para la distribución de Weibull. [*Sugerencia*: En la definición de una fda haga la transformación $z = y^{\beta}$].

6.84 Explique por qué la naturaleza del escenario en el ejercicio de repaso 6.82 probablemente no se preste a la distribución exponencial.

6.85 A partir de la relación entre la variable aleatoria chi cuadrada y la variable aleatoria gamma, demuestre que la media de la variable aleatoria chi cuadrada es v y que la varianza es $2v$.

6.86 El tiempo que le toma a un usuario de computadora leer su correo electrónico, en segundos, se distribuye como una variable aleatoria logarítmica normal con $\mu = 1.8$ y $\sigma^2 = 4.0$.
a) ¿Cuál es la probabilidad de que el usuario lea el correo durante más de 20 segundos? ¿Y por más de un minuto?
b) ¿Cuál es la probabilidad de que el usuario lea el correo durante un tiempo que sea igual a la media de la distribución logarítmica normal subyacente?

6.87 **Proyecto de grupo:** Pida a grupos de estudiantes que observen durante 2 semanas el número de personas que entra a una cafetería o restaurante de comida rápida específico en el transcurso de una hora, empezando a la misma hora cada día. La hora deberá ser la de mayor tránsito en la cafetería o restaurante. Los datos reunidos corresponderán al número de clientes que entran al lugar durante cada lapso de media hora. De esta manera, cada día se recolectarán 2 datos. Supongamos que la variable aleatoria X, el número de personas que entra cada media hora, tiene una distribución de Poisson. Los estudiantes deberán calcular la media y la varianza muestrales de X utilizando los 28 datos obtenidos.
a) ¿Qué evidencia hay de que la distribución de Poisson es o no correcta?
b) Dado que X es una variable de Poisson, ¿cuál es la distribución de T, el tiempo entre la llegada de las personas al lugar durante un lapso de media hora? Proporcione un estimado numérico del parámetro de esa distribución.
c) Proporcione un estimado de la probabilidad de que el lapso de tiempo entre las 2 llegadas sea menor de 15 minutos.
d) ¿Cuál es la probabilidad estimada de que el lapso entre las 2 llegadas sea mayor de 10 minutos?
e) ¿Cuál es la probabilidad estimada de que 20 minutos después de iniciar la recolección de datos ningún cliente haya llegado?

6.11 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Muchos de los riesgos en el uso del material de este capítulo son muy similares a los del capítulo 5. Uno de los peores abusos de la estadística consiste en suponer que se trata de

una distribución normal haciendo algún tipo de inferencia estadística, cuando en realidad no es normal. En los capítulos 10 al 15 el lector estudiará las pruebas de hipótesis, en las que se asume normalidad. Además, se le recordará al lector que hay **pruebas de la bondad de ajuste**, además de las rutinas gráficas que se examinan en los capítulos 8 y 10, que permiten verificar los datos para determinar si es razonable la suposición de normalidad.

Debemos hacer advertencias similares con respecto a las suposiciones que a menudo se hacen sobre otras distribuciones, además de la curva normal. En este libro se han presentado ejemplos en los que es necesario calcular las probabilidades de falla de ciertos productos o la probabilidad de recibir una queja durante cierto periodo. Se suelen hacer suposiciones con respecto a cierto tipo de distribución, así como a los valores de los parámetros de la distribución. Observe que los problemas de ejemplo incluyen los valores de los parámetros (por ejemplo, el valor de β para la distribución exponencial). No obstante, en los problemas de la vida real los valores de los parámetros deben ser estimaciones de experiencias o datos reales. Observe el énfasis que se pone en la estimación en los proyectos que aparecen en los capítulos 1, 5 y 6, así como la referencia que se hace en el capítulo 5 a las estimación de parámetros, tema que se analizará ampliamente a partir del capítulo 9.

Capítulo 7

Funciones de variables aleatorias (opcional)

7.1 Introducción

Este capítulo contiene un amplio espectro de material. Los capítulos 5 y 6 tratan tipos específicos de distribuciones, tanto discretas como continuas. Éstas son distribuciones que suelen aplicarse en muchos campos, por ejemplo en el de la confiabilidad, el de control de calidad y el de muestreo de aceptación. En este capítulo comenzamos a estudiar un tema más general: el de la distribución de funciones de variables aleatorias. Se presentan las técnicas generales y se ilustran con ejemplos. Las presentaciones van seguidas por un concepto relacionado, el de *funciones generadoras de momentos*, que pueden ser útiles para el aprendizaje de distribuciones de funciones lineales de variables aleatorias.

En los métodos estadísticos estándar, el resultado de la prueba de hipótesis estadísticas, la estimación, o incluso las gráficas estadísticas, no involucra a una sola variable aleatoria sino a *funciones de una o más variables aleatorias*. Como resultado, la inferencia estadística requiere la distribución de tales funciones. Por ejemplo, es común que se utilicen **promedios de variables aleatorias**. Además, las sumatorias y las combinaciones lineales más generales son importantes. Con frecuencia nos interesa la distribución de las sumas de cuadrados de variables aleatorias, en particular la manera en que se utilizan las técnicas del análisis de varianza, las cuales se estudiarán en los capítulos 11 a 14.

7.2 Transformaciones de variables

Con frecuencia, en la estadística se enfrenta la necesidad de derivar la distribución de probabilidad de una función de una o más variables aleatorias. Por ejemplo, suponga que X es una variable aleatoria discreta con distribución de probabilidad $f(x)$, suponga también que $Y = u\,(X)$ define una transformación uno a uno entre los valores de X y Y. Queremos encontrar la distribución de probabilidad de Y. Es importante notar que la transformación uno a uno implica que cada valor x está relacionado con un, y sólo un, valor $y = u(x)$, y que cada valor y está relacionado con un, y sólo un, valor $x = w(y)$, donde $w(y)$ se obtiene al resolver $y = u(x)$ para x en términos de y.

A partir de lo expuesto respecto a las distribuciones de probabilidad discreta en el capítulo 3, nos quedó claro que la variable aleatoria Y toma el valor y cuando X toma el valor $w(y)$. En consecuencia, la distribución de probabilidad de Y es dada por

$$g(y) = P(Y = y) = P[X = w(y)] = f[w(y)].$$

Teorema 7.1: Suponga que X es una variable aleatoria **discreta** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva exclusivamente para x en términos de y, digamos, $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)].$$

Ejemplo 7.1: Sea X una variable aleatoria geométrica con la siguiente distribución de probabilidad

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \qquad x = 1, 2, 3, \ldots.$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

***Solución*:** Como todos los valores de X son positivos, la transformación define una correspondencia uno a uno entre los valores x y y, $y = x^2$ y $x = \sqrt{y}$. Por lo tanto,

$$g(y) = \begin{cases} f(\sqrt{y}) = \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 4, 9, \ldots, \\ 0, & \text{en cualquier caso.} \end{cases}$$

De manera similar, para una transformación de dos dimensiones, tenemos el resultado en el teorema 7.2.

Teorema 7.2: Suponga que X_1 y X_2 son variables aleatorias **discretas**, con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones

$$y_1 = u_1(x_1, x_2) \qquad \text{y} \qquad y_2 = u_2(x_1, x_2)$$

se pueden resolver exclusivamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

El teorema 7.2 es muy útil para encontrar la distribución de alguna variable aleatoria $Y_1 = u_1(X_1, X_2)$, donde X_1 y X_2 son variables aleatorias discretas con distribución de probabilidad conjunta $f(x_1, x_2)$. Definimos simplemente una segunda función, digamos $Y_2 = u_2(X_1, X_2)$, manteniendo una correspondencia uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), y obtenemos la distribución de probabilidad conjunta $g(y_1, y_2)$. La distribución de Y_1 es precisamente la distribución marginal de $g(y_1, y_2)$ que se encuentra sumando los valores y_2. Si denotamos la distribución de Y_1 con $h(y_1)$, podemos escribir

$$h(y_1) = \sum_{y_2} g(y_1, y_2).$$

Ejemplo 7.2: Sean X_1 y X_2 dos variables aleatorias independientes que tienen distribuciones de Poisson con los parámetros μ_1 y μ_2, respectivamente. Calcule la distribución de la variable aleatoria $Y_1 = X_1 + X_2$.

***Solución*:** Como X_1 y X_2 son independientes, podemos escribir

$$f(x_1, x_2) = f(x_1)f(x_2) = \frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\frac{e^{-\mu_2}\mu_2^{x_2}}{x_2!} = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{x_1}\mu_2^{x_2}}{x_1!x_2!},$$

donde $x_1 = 0, 1, 2,...$ y $x_2 = 0, 1, 2,....$ Definamos ahora una segunda variable aleatoria, digamos $Y_2 = X_2$. Las funciones inversas son dadas por $x_1 = y_1 - y_2$ y $x_2 = y_2$. Si usamos el teorema 7.2, encontramos que la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!},$$

donde $y_1 = 0, 1, 2,...$ y $y_2 = 0, 1, 2,..., y_1$. Advierta que, como $x_1 > 0$, la transformación $x_1 = y_1 - x_2$ implica que y_2 y, por lo tanto, x_2 siempre deben ser menores o iguales que y_1. En consecuencia, la distribución de probabilidad marginal de Y_1 es

$$\begin{aligned} h(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) = e^{-(\mu_1+\mu_2)} \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{y_2!(y_1 - y_2)!}\mu_1^{y_1-y_2}\mu_2^{y_2} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \binom{y_1}{y_2}\mu_1^{y_1-y_2}\mu_2^{y_2}. \end{aligned}$$

Al reconocer esta suma como la expansión binomial de $(\mu_1 + \mu_2)^{y1}$, obtenemos

$$h(y_1) = \frac{e^{-(\mu_1+\mu_2)}(\mu_1 + \mu_2)^{y_1}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots,$$

a partir de lo cual concluimos que la suma de las dos variables aleatorias independientes que tienen distribuciones de Poisson, con los parámetros μ_1 y μ_2, tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$. ∎

Para calcular la distribución de probabilidad de la variable aleatoria $Y = u(X)$, cuando X es una variable aleatoria continua y la transformación es uno a uno, necesitaremos el teorema 7.3. La demostración de este teorema se deja al lector.

Teorema 7.3: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una correspondencia uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva exclusivamente para x en términos de y, digamos $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)]|J|,$$

donde $J = w'(y)$ y se llama **jacobiano** de la transformación.

Ejemplo 7.3: Sea X una variable aleatoria continua con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{x}{12}, & 1 < x < 5, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = 2X - 3$.

***Solución*:** La solución inversa de $y = 2x - 3$ produce $x = (y + 3)/2$, de la que obtenemos $J = w'(y) = dx/dy = 1/2$. Por lo tanto, usando el teorema 7.3 encontramos que la función de densidad de Y es

$$g(y) = \begin{cases} \frac{(y+3)/2}{12}\left(\frac{1}{2}\right) = \frac{y+3}{48}, & -1 < y < 7, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Para calcular la distribución de probabilidad conjunta de las variables aleatorias $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$, cuando X_1 y X_2 son continuas y la transformación es uno a uno, necesitamos un teorema adicional análogo al teorema 7.2, el cual establecemos sin demostración.

Teorema 7.4: Suponga que X_1 y X_2 son variables aleatorias **continuas** con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones $y_1 = u_1(x_1, x_2)$ y $y_2 = u_2(x_1, x_2)$ se resuelven exclusivamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

donde el jacobiano es el determinante 2×2

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

y $\frac{\partial x_1}{\partial y_1}$ es simplemente la derivada de $x_1 = w_1(y_1, y_2)$ respecto a y_1, con y_2 constante, que en cálculo se denomina derivada parcial de x_1 respecto a y_1. Las otras derivadas parciales se definen de manera similar.

Ejemplo 7.4: Sean X_1 y X_2 dos variables aleatorias continuas con la siguiente distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} 4x_1x_2, & 0 < x_1 < 1, \ 0 < x_2 < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad conjunta de $Y_1 = X_1^2$ y $Y_2 = X_1X_2$.

***Solución*:** Las soluciones inversas de $y_1 = x_1^2$ y $y_2 = x_1x_2$ son $x_1 = \sqrt{y_1}$ y $x_2 = y_2/\sqrt{y_1}$, de las que obtenemos

$$J = \begin{vmatrix} 1/(2\sqrt{y_1}) & 0 \\ -y_2/2y_1^{3/2} & 1/\sqrt{y_1} \end{vmatrix} = \frac{1}{2y_1}.$$

Para determinar el conjunto B de puntos en el plano y_1y_2 en el que se traza el conjunto A de puntos en el plano x_1x_2 escribimos

$$x_1 = \sqrt{y_1} \qquad \text{y} \qquad x_2 = y_2/\sqrt{y_1}.$$

Luego, al establecer $x_1 = 0$, $x_2 = 0$, $x_1 = 1$ y $x_2 = 1$, las fronteras del conjunto A se transforman en $y_1 = 0$, $y_2 = 0$, $y_1 = 1$ y $y_2 = \sqrt{y_1}$ o $y_2^2 = y_1$. Las dos regiones se ilustran en la figura 7.1. Al trazar el conjunto $A = \{(x_1, x_2) \mid 0 < x_1 < 1,\ 0 < x_2 < 1\}$ en el conjunto $B = \{(y_1, y_2) \mid y_2^2 < y_1 < 1,\ 0 < y_2 < 1\}$, se vuelve evidente que la transformación es uno a uno. Del teorema 7.4, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = 4(\sqrt{y_1})\frac{y_2}{\sqrt{y_1}}\frac{1}{2y_1} = \begin{cases} \frac{2y_2}{y_1}, & y_2^2 < y_1 < 1,\ 0 < y_2 < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

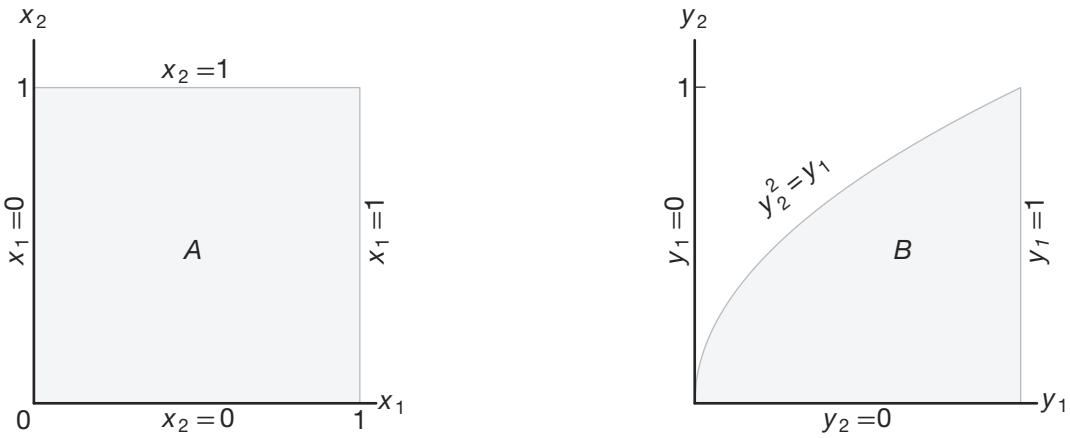


Figura 7.1: Gráfica del conjunto A en el conjunto B.

A menudo surgen problemas cuando deseamos encontrar la distribución de probabilidad de la variable aleatoria $Y = u(X)$ y X es una variable aleatoria continua y la transformación no es uno a uno. Es decir, a cada valor x le corresponde exactamente un valor y; pero a cada valor y le corresponde más de un valor x. Por ejemplo, suponga que $f(x)$ es positiva en el intervalo $-1 < x < 2$ y cero en cualquier caso. Considere la transformación $y = x^2$. En este caso, $x = \pm\sqrt{y}$ para $0 < y < 1$ y $x = \sqrt{y}$ para $1 < y < 4$. Para el intervalo $1 < y < 4$, la distribución de probabilidad de Y se calcula como antes, con el teorema 7.3. Es decir,

$$g(y) = f\,[w(y)]|J\,| = \frac{f\,(\sqrt{y})}{2\sqrt{y}}, \qquad 1 < y < 4.$$

Sin embargo, cuando $0 < y < 1$, podemos dividir el intervalo $-1 < x < 1$ para obtener las dos funciones inversas

$$x = -\sqrt{y}, \qquad -1 < x < 0, \qquad \text{y} \qquad x = \sqrt{y}, \qquad 0 < x < 1.$$

Entonces, a todo valor y le corresponde un solo valor x para cada partición. En la figura 7.2 vemos que

$$P(a < Y < b) = P(-\sqrt{b} < X < -\sqrt{a}) + P(\sqrt{a} < X < \sqrt{b})$$
$$= \int_{-\sqrt{b}}^{-\sqrt{a}} f(x)\, dx + \int_{\sqrt{a}}^{\sqrt{b}} f(x)\, dx.$$

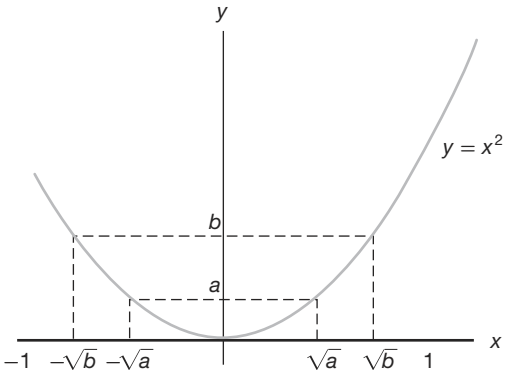


Figura 7.2: Función decreciente y creciente.

Al cambiar la variable de integración de x a y, obtenemos

$$P(a < Y < b) = \int_b^a f(-\sqrt{y})J_1\, dy + \int_a^b f(\sqrt{y})J_2\, dy$$
$$= -\int_a^b f(-\sqrt{y})J_1\, dy + \int_a^b f(\sqrt{y})J_2\, dy,$$

donde

$$J_1 = \frac{d(-\sqrt{y})}{dy} = \frac{-1}{2\sqrt{y}} = -|J_1|$$

y

$$J_2 = \frac{d(\sqrt{y})}{dy} = \frac{1}{2\sqrt{y}} = |J_2|.$$

Por lo tanto, podemos escribir

$$P(a < Y < b) = \int_a^b [f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2|]\, dy,$$

y entonces

$$g(y) = f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2| = \frac{f(-\sqrt{y}) + f(\sqrt{y})}{2\sqrt{y}}, \qquad 0 < y < 1.$$

La distribución de probabilidad de Y para $0 < y < 4$ se puede escribir ahora como

$$g(y) = \begin{cases} \frac{f(-\sqrt{y})+f(\sqrt{y})}{2\sqrt{y}}, & 0 < y < 1, \\ \frac{f(\sqrt{y})}{2\sqrt{y}}, & 1 < y < 4, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Este procedimiento para calcular $g(y)$ cuando $0 < y < 1$ se generaliza en el teorema 7.5 para k funciones inversas. Para transformaciones de funciones de diversas variables que no son uno a uno se recomienda al lector *Introduction to Mathematical Statistics* de Hogg, McKean y Craig (2005; véase la bibliografía).

Teorema 7.5: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación entre los valores de X y Y que no es uno a uno. Si el intervalo sobre el que se define X se puede dividir en k conjuntos mutuamente disjuntos de manera que cada una de las funciones inversas

$$x_1 = w_1(y), \quad x_2 = w_2(y), \quad \ldots, \quad x_k = w_k(y)$$

de $y = u(x)$ defina una correspondencia uno a uno, entonces la distribución de probabilidad de Y es

$$g(y) = \sum_{i=1}^{k} f[w_i(y)]|J_i|,$$

donde $J_i = w_i'(y)$, $i = 1, 2, \ldots, k$.

Ejemplo 7.5: Demuestre que $Y = (X - \mu)^2/\sigma^2$ tiene una distribución chi cuadrada con 1 grado de libertad cuando X tiene una distribución normal con media μ y varianza σ^2.

***Solución*:** Sea $Z = (X - \mu)/\sigma$, donde la variable aleatoria Z tiene la distribución normal estándar

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

Ahora debemos calcular la distribución de la variable aleatoria $Y = Z^2$. Las soluciones inversas de $y = z^2$ son $z = \pm\sqrt{y}$. Si designamos $z_1 = -\sqrt{y}$ y $z_2 = \sqrt{y}$, entonces $J_1 = -1/2\sqrt{y}$ y $J_2 = -1/2\sqrt{y}$. Entonces, por el teorema 7.5, tenemos

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{-1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} y^{1/2-1} e^{-y/2}, \qquad y > 0.$$

Como $g(y)$ es una función de densidad, se deduce que

$$1 = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{1/2-1} e^{-y/2}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}} \int_0^\infty \frac{y^{1/2-1} e^{-y/2}}{\sqrt{2}\Gamma(1/2)}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}},$$

la integral es el área bajo una curva de probabilidad gamma con los parámetros $\alpha = 1/2$ y $\beta = 2$. Por lo tanto, $\sqrt{\pi} = \Gamma(1/2)$ y la densidad de Y es dada por

$$g(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, & y > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

que se considera una distribución chi cuadrada con 1 grado de libertad. ∎

7.3 Momentos y funciones generadoras de momentos

En esta sección nos concentramos en aplicaciones de las funciones generadoras de momentos. El propósito evidente de la función generadora de momentos es la determinación de los momentos de variables aleatorias. Sin embargo, la contribución más importante consiste en establecer distribuciones de funciones de variables aleatorias.

Si $g(X) = X^r$ para $r = 0, 1, 2, 3,...$, la definición 7.1 proporciona un valor esperado que se denomina *r*-ésimo **momento alrededor del origen** de la variable aleatoria X, que denotamos con μ'_r.

Definición 7.1: El *r*-ésimo **momento alrededor del origen** de la variable aleatoria X es dado por

$$\mu'_r = E(X^r) = \begin{cases} \sum_x x^r f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} x^r f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Como el primer y segundo momentos alrededor del origen son dados por $\mu'_1 = E(X)$ y $\mu'_2 = E(X^2)$, podemos escribir la media y la varianza de una variable aleatoria como

$$\mu = \mu'_1 \qquad \text{y} \qquad \sigma^2 = \mu'_2 - \mu^2.$$

Aunque los momentos de una variable aleatoria se pueden determinar directamente a partir de la definición 7.1, existe un procedimiento alternativo, el cual requiere que utilicemos una **función generadora de momentos**.

Definición 7.2: La **función generadora de momentos** de la variable aleatoria X es dada por $E(e^{tX})$, y se denota con $M_X(t)$. Por lo tanto,

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Las funciones generadoras de momentos existirán sólo si la sumatoria o integral de la definición 7.2 converge. Si existe una función generadora de momentos de una variable aleatoria X, se puede utilizar para generar todos los momentos de dicha variable. El método se describe en el teorema 7.6 sin demostración.

Teorema 7.6: Sea X una variable aleatoria con función generadora de momentos $M_X(t)$. Entonces,

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \mu'_r.$$

Ejemplo 7.6: Calcule la función generadora de momentos de la variable aleatoria binomial X y después utilícela para verificar que $\mu = np$ y $\sigma^2 = npq$.

***Solución*:** A partir de la definición 7.2 tenemos

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x}.$$

Al reconocer a esta última sumatoria como la expansión binomial de $(pe^t + q)^n$ obtenemos

$$M_X(t) = (pe^t + q)^n.$$

Así,

$$\frac{dM_X(t)}{dt} = n(pe^t + q)^{n-1}pe^t$$

y

$$\frac{d^2M_X(t)}{dt^2} = np[e^t(n-1)(pe^t + q)^{n-2}pe^t + (pe^t + q)^{n-1}e^t].$$

Al establecer $t = 0$ obtenemos

$$\mu_1' = np \text{ y } \mu_2' = np[(n-1)p + 1].$$

Por consiguiente,

$$\mu = \mu_1' = np \text{ y } \sigma^2 = \mu_2' - \mu^2 = np(1-p) = npq,$$

que coincide con los resultados que se obtuvieron en el capítulo 5. ∎

Ejemplo 7.7: Demuestre que la función generadora de momentos de la variable aleatoria X, la cual tiene una distribución de probabilidad normal con media μ y varianza σ^2, es dada por

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

***Solución*:** A partir de la definición 7.2, la función generadora de momentos de la variable aleatoria normal X es

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{x^2 - 2(\mu + t\sigma^2)x + \mu^2}{2\sigma^2}\right] dx. \end{aligned}$$

Si completamos el cuadrado en el exponente, podemos escribir

$$x^2 - 2(\mu + t\sigma^2)x + \mu^2 = [x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4$$

y, entonces,

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4}{2\sigma^2}\right\} dx \\ &= \exp\left(\frac{2\mu t + \sigma^2 t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2}{2\sigma^2}\right\} dx. \end{aligned}$$

Sea $w = [x - (\mu + t\sigma^2)]/\sigma$; entonces $dx = \sigma\, dw$ y

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

ya que la última integral representa el área bajo una curva de densidad normal estándar y, en consecuencia, es igual a 1. ∎

Aunque el método de transformación de variables brinda una forma eficaz para determinar la distribución de una función de múltiples variables, existe un procedimiento alternativo, y que a menudo se prefiere cuando la función a analizar es una combinación lineal de variables aleatorias independientes. Este procedimiento utiliza las propiedades de las funciones generadoras de momentos que se estudian en los siguientes cuatro teoremas. Para no rebasar el alcance matemático de este libro, establecemos el teorema 7.7 sin demostración.

Teorema 7.7: (**Teorema de unicidad**) Sean X y Y dos variables aleatorias con funciones generadoras de momentos $M_X(t)$ y $M_Y(t)$, respectivamente. Si $M_X(t) = M_Y(t)$ para todos los valores de t, entonces X y Y tienen la misma distribución de probabilidad.

Teorema 7.8: $M_{X+a}(t) = e^{at} M_X(t)$.

***Prueba*:** $M_{X+a}(t) = E[e^{t(X+a)}] = e^{at} E(e^{tX}) = e^{at} M_X(t)$. ∎

Teorema 7.9: $M_{aX}(t) = M_X(at)$.

***Prueba*:** $M_{aX}(t) = E[e^{t(aX)}] = E[e^{(at)X}] = M_X(at)$. ∎

Teorema 7.10: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con funciones generadoras de momentos $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectivamente, y $Y = X_1 + X_2 + \ldots + X_n$, entonces,

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t).$$

La demostración del teorema 7.10 se deja al lector.

Los teoremas 7.7 a 7.10 son fundamentales para entender las funciones generadoras de momentos. A continuación se presenta un ejemplo como ilustración. Hay muchas situaciones en que necesitamos conocer la distribución de la suma de las variables aleatorias. Podemos utilizar los teoremas 7.7 y 7.10, así como el resultado del ejercicio 7.19 de la página 224, para calcular la distribución de una suma de dos variables aleatorias independientes de Poisson, con funciones generadoras de momentos dadas por

$$M_{X_1}(t) = e^{\mu_1(e^t - 1)} \text{ y } M_{X_2}(t) = e^{\mu_2(e^t - 1)},$$

respectivamente. De acuerdo con el teorema 7.10, la función generadora de momentos de la variable aleatoria $Y_1 = X_1 + X_2$ es

$$M_{Y_1}(t) = M_{X_1}(t) M_{X_2}(t) = e^{\mu_1(e^t - 1)} e^{\mu_2(e^t - 1)} = e^{(\mu_1 + \mu_2)(e^t - 1)},$$

que de inmediato identificamos como la función generadora de momentos de una variable aleatoria que tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$. Por lo tanto, de acuerdo con el teorema 7.7, de nuevo concluimos que la suma de dos variables aleatorias independientes, que tienen distribuciones de Poisson con los parámetros μ_1 y μ_2, tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$.

Combinaciones lineales de variables aleatorias

En estadística aplicada a menudo se necesita conocer la distribución de probabilidad de una combinación lineal de variables aleatorias normales independientes. Obtengamos la distribución de la variable aleatoria $Y = a_1X_1 + a_2X_2$ cuando X_1 es una variable normal con media μ_1 y varianza σ_1^2 y X_2 también es una variable normal, pero independiente de X_1, con media μ_2 y varianza σ_2^2. Primero, por medio del teorema 7.10, obtenemos

$$M_Y(t) = M_{a_1X_1}(t)M_{a_2X_2}(t),$$

y después, usando el teorema 7.9, obtenemos

$$M_Y(t) = M_{X_1}(a_1t)M_{X_2}(a_2t).$$

Si sustituimos a_1t por t, y despues a_2t por t, en una función generadora de momentos de la distribución normal derivada en el ejemplo 7.7, tenemos

$$\begin{aligned} M_Y(t) &= \exp(a_1\mu_1t + a_1^2\sigma_1^2t^2/2 + a_2\mu_2t + a_2^2\sigma_2^2t^2/2) \\ &= \exp[(a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2], \end{aligned}$$

que reconocemos como la función generadora de momentos de una distribución que es normal, con media $a_1\mu_1 + a_2\mu_2$ y varianza $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Al generalizar para el caso de n variables normales independientes, establecemos el siguiente resultado.

Teorema 7.11: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \ldots \mu_n$ y varianzas $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectivamente, entonces la variable aleatoria

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

tiene una distribución normal con media

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

y varianza

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

Ahora es evidente que la distribución de Poisson y la distribución normal tienen una propiedad reproductiva, en el sentido de que la suma de variables aleatorias independientes que tengan cualquiera de estas distribuciones es una variable aleatoria que también tiene el mismo tipo de distribución. La distribución chi cuadrada también posee esta propiedad reproductiva.

Teorema 7.12: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias mutuamente independientes, que tienen distribuciones chi cuadrada con $v_1, v_2, \ldots, v_n$ grados de libertad, respectivamente, entonces la variable aleatoria

$$Y = X_1 + X_2 + \cdots + X_n$$

tiene una distribución chi cuadrada con $v = v_1 + v_2 + \ldots + v_n$ grados de libertad.

***Prueba*:** Por medio del teorema 7.10 y el ejercicio 7.21,

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) \text{ y } M_{X_i}(t) = (1-2t)^{-v_i/2},\ i = 1, 2, \ldots, n.$$

Por lo tanto,

$$M_Y(t) = (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} \ldots (1-2t)^{-v_n/2} = (1-2t)^{-(v_1+v_2+\cdots+v_n)/2},$$

que reconocemos como la función generadora de momentos de una distribución chi cuadrada con $v = v_1 + v_2 + \ldots + v_n$ grados de libertad. ∎

Corolario 7.1: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales idénticas, con media μ y varianza σ^2, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Este corolario es una consecuencia inmediata del ejemplo 7.5, y establece una relación entre la muy importante distribución chi cuadrada y la distribución normal. También debe brindar al lector una idea muy clara de lo que significa el parámetro llamado grados de libertad. En futuros capítulos el concepto de grados de libertad desempeñará un papel cada vez más relevante.

Corolario 7.2: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y X_i tiene una distribución normal con media μ_i y varianza σ_i^2 para $i = 1, 2, \ldots, n$, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Ejercicios

7.1 Sea X una variable aleatoria que tiene la siguiente probabilidad

$$f(x) = \begin{cases} \frac{1}{3}, & x = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = 2X - 1$.

7.2 Sea X una variable aleatoria binomial con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \binom{3}{x} \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{3-x}, & x = 0, 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.3 Sean X_1 y X_2 variables aleatorias discretas con la siguiente distribución multinomial conjunta

$$f(x_1, x_2) = \binom{2}{x_1, x_2, 2 - x_1 - x_2} \left(\frac{1}{4}\right)^{x_1} \left(\frac{1}{3}\right)^{x_2} \left(\frac{5}{12}\right)^{2-x_1-x_2}$$

para $x_1 = 0, 1, 2$; $x_2 = 0, 1, 2$; $x_1 + x_2 \le 2$; y cero en cualquier caso. Calcule la distribución de probabilidad conjunta de $Y_1 = X_1 + X_2$ y $Y_2 = X_1 - X_2$.

7.4 Sean X_1 y X_2 variables aleatorias discretas con la siguiente distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{18}, & x_1 = 1, 2; \ x_2 = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X_1 X_2$.

7.5 Si X tiene la siguiente distribución de probabilidad

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que la variable aleatoria $Y = -2\ln X$ tiene una distribución chi cuadrada con 2 grados de libertad.

7.6 Dada la variable aleatoria X con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

calcule la distribución de probabilidad de $Y = 8X^3$.

7.7 La velocidad de una molécula en un gas uniforme en equilibrio es una variable aleatoria V, cuya distribución de probabilidad es dada por

$$f(v) = \begin{cases} kv^2 e^{-bv^2}, & v > 0, \\ 0, & \text{en cualquier caso,} \end{cases}$$

donde k es una constante adecuada y b depende de la temperatura absoluta y de la masa de la molécula. Calcule la distribución de probabilidad de la energía cinética de la molécula W, donde $W = mV^2/2$.

7.8 La utilidad de un distribuidor, en unidades de \$5000, sobre un automóvil nuevo, es dada por $Y = X^2$, donde X es una variable aleatoria que tiene la siguiente función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria Y.
b) Utilice la función de densidad de Y para calcular la probabilidad de que la utilidad sobre el siguiente automóvil nuevo que venda este distribuidor sea menor que \$500.

7.9 El periodo hospitalario, en días, para pacientes que siguen un tratamiento para cierto tipo de enfermedad del riñón es una variable aleatoria $Y = X + 4$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria Y.
b) Utilice la función de densidad de Y para calcular la probabilidad de que el periodo hospitalario para un paciente que sigue este tratamiento exceda los 8 días.

7.10 Las variables aleatorias X y Y, que representan los pesos de cremas y chiclosos, respectivamente, en cajas de un kilogramo de chocolates que contienen una combinación de cremas, chiclosos y envinados, tienen la siguiente función de densidad conjunta

$$f(x, y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, x + y \le 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria $Z = X + Y$.
b) Utilice la función de densidad de Z para calcular la probabilidad de que, en una determinada caja, la suma de los pesos de las cremas y los chiclosos sea por lo menos 1/2 del peso total, pero menos de 3/4.

7.11 La cantidad de queroseno en un tanque al inicio de cualquier día, en miles de litros, es una cantidad aleatoria Y, de la cual una cantidad aleatoria X se vende durante ese día. Suponga que la función de densidad conjunta de estas variables es dada por

$$f(x, y) = \begin{cases} 2, & 0 < x < y,\ 0 < y < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la función de densidad de probabilidad para la cantidad de queroseno que queda en el tanque al final del día.

7.12 Sean X_1 y X_2 variables aleatorias independientes que tienen cada una la siguiente distribución de probabilidad

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que las variables aleatorias Y_1 y Y_2 son independientes cuando $Y_1 = X_1 + X_2$ y $Y_2 = X_1/(X_1 + X_2)$.

7.13 Una corriente de I amperios que fluye a través de una resistencia de R ohms varía de acuerdo con la siguiente distribución de probabilidad

$$f(i) = \begin{cases} 6i(1-i), & 0 < i < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Si la resistencia varía independientemente de la corriente de acuerdo con la siguiente distribución de probabilidad

$$g(r) = \begin{cases} 2r, & 0 < r < 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

calcule la distribución de probabilidad para la potencia $W = I^2R$ watts.

7.14 Sea X una variable aleatoria con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{1+x}{2}, & -1 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.15 Si X tiene la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{2(x+1)}{9}, & -1 < x < 2, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.16 Demuestre que el r-ésimo momento respecto al origen de la distribución gamma es

$$\mu'_r = \frac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}.$$

[*Sugerencia*: Sustituya $y = x/\beta$ en la integral que define μ'_r y después utilice la función gamma para evaluar la integral].

7.17 Una variable aleatoria X tiene la siguiente distribución uniforme discreta

$$f(x; k) = \begin{cases} \frac{1}{k}, & x = 1, 2, \ldots, k, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que la función generadora de momentos de X es

$$M_X(t) = \frac{e^t(1 - e^{kt})}{k(1 - e^t)}.$$

7.18 Una variable aleatoria X tiene la distribución geométrica $g(x; p) = pq^{x-1}$ para $x = 1, 2, 3, \ldots$. Demuestre que la función generadora de momentos de X es

$$M_X(t) = \frac{pe^t}{1 - qe^t}, \qquad t < \ln q,$$

y después use $M_X(t)$ para calcular la media y la varianza de la distribución geométrica.

7.19 Una variable aleatoria X tiene la distribución de Poisson $p(x; \mu) = e^{-\mu}\mu^x/x!$ para $x = 0, 1, 2, \ldots$. Demuestre que la función generadora de momentos de X es

$$M_X(t) = e^{\mu(e^t - 1)}.$$

Utilice $M_X(t)$ para calcular la media y la varianza de la distribución de Poisson.

7.20 La función generadora de momentos de cierta variable aleatoria de Poisson X es dada por

$$M_X(t) = e^{4(e^t - 1)}.$$

Calcule $P(\mu - 2\sigma < X < \mu + 2\sigma)$.

7.21 Demuestre que la función generadora de momentos de la variable aleatoria X, que tiene una distribución chi cuadrada con v grados de libertad, es

$$M_X(t) = (1 - 2t)^{-v/2}.$$

7.22 Con la función generadora de momentos del ejemplo 7.21 demuestre que la media y la varianza de la distribución chi cuadrada con v grados de libertad son, respectivamente, v y $2v$.

7.23 Si tanto X como Y, distribuidas de manera independiente, siguen distribuciones exponenciales con parámetro medio 1, calcule las distribuciones de

a) $U = X + Y$;
b) $V = X/(X + Y)$.

7.24 Mediante la expansión de e^{tx} en una serie de Maclaurin y la integración término por término, demuestre que

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$
$$= 1 + \mu t + \mu'_2 \frac{t^2}{2!} + \ldots + \mu'_r \frac{t^r}{r!} + \ldots$$

Capítulo 8

Distribuciones de muestreo fundamentales y descripciones de datos

8.1 Muestreo aleatorio

El resultado de un experimento estadístico se puede registrar como un valor numérico o como una representación descriptiva. Cuando se lanza un par de dados y lo que nos interesa es el resultado total, registramos un valor numérico. Sin embargo, si a los estudiantes de cierta escuela se les hacen pruebas de sangre para averiguar cuál es su tipo, podría ser más útil una representación descriptiva. La sangre de una persona se puede clasificar de 8 maneras. Puede ser AB, A, B u O, cada una con un signo de más o de menos, lo cual depende de la presencia o ausencia del antígeno Rh.

En este capítulo nos enfocamos en el muestreo de distribuciones o poblaciones, y estudiamos cantidades tan importantes como la *media de la muestra* y la *varianza de la muestra*, que serán de importancia fundamental en los capítulos siguientes. Además, en los próximos capítulos intentamos introducir al lector al papel que desempeñarán la media y la varianza de la muestra en la inferencia estadística. El uso de las computadoras modernas de alta velocidad permite a los científicos e ingenieros incrementar enormemente su uso de la inferencia estadística formal con técnicas gráficas. La mayoría de las veces la inferencia formal parece muy árida y quizás incluso abstracta para el profesional o el gerente que desea que el análisis estadístico sea una guía para la toma de decisiones.

Poblaciones y muestras

Comenzamos esta sección presentando los conceptos de *poblaciones* y *muestras*. Ambas se mencionan de forma extensa en el capítulo 1; sin embargo, aquí será necesario estudiarlas más ampliamente, en particular en el contexto del concepto de variables aleatorias. La totalidad de observaciones que nos interesan, ya sean de número finito o infinito, constituye lo que llamamos **población**. En alguna época el término *población* se refería a observaciones que se obtenían de estudios estadísticos aplicados a personas. En la actualidad el estadístico utiliza la palabra para referirse a observaciones sobre cualquier cuestión de interés, ya sea de grupos de personas, de animales o de todos los resultados posibles de algún complicado sistema biológico o de ingeniería.

Definición 8.1: Una **población** consta de la totalidad de las observaciones en las que estamos interesados.

El número de observaciones en la población se define como el tamaño de la población. Si en la escuela hay 600 estudiantes que clasificamos de acuerdo con su tipo de sangre, decimos que tenemos una población de tamaño 600. Los números en las cartas de una baraja, las estaturas de los residentes de cierta ciudad y las longitudes de los peces en un lago específico son ejemplos de poblaciones de tamaño finito. En cada caso el número total de observaciones es un número finito. Las observaciones que se obtienen al medir diariamente la presión atmosférica desde el pasado hasta el futuro, o todas las mediciones de la profundidad de un lago desde cualquier posición concebible son ejemplos de poblaciones cuyos tamaños son infinitos. Algunas poblaciones finitas son tan grandes que en teoría las supondríamos infinitas, lo cual es cierto si se considera la población de la vida útil de cierto tipo de batería de almacenamiento que se está fabricando para distribuirla en forma masiva en todo el país.

Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$. Si se inspeccionan artículos que salen de una línea de ensamble para buscar defectos, entonces cada observación en la población podría ser un valor 0 o 1 de la variable aleatoria X de Bernoulli, con una distribución de probabilidad

$$b(x; 1, p) = p^x q^{1-x}, \qquad x = 0, 1$$

donde 0 indica un artículo sin defecto y 1 indica un artículo defectuoso. De hecho, se supone que p, la probabilidad de que cualquier artículo esté defectuoso, permanece constante de una prueba a otra. En el experimento del tipo de sangre la variable aleatoria X representa el tipo de sangre y se supone que toma un valor del 1 al 8. A cada estudiante se le asigna uno de los valores de la variable aleatoria discreta. Las duraciones de las baterías de almacenamiento son valores que toma una variable aleatoria continua que quizá tiene una distribución normal. De ahora en adelante, cuando nos refiramos a una "población binomial", a una "población normal" o, en general, a la "población $f(x)$", aludiremos a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o la distribución de probabilidad $f(x)$. Por ello, a la media y a la varianza de una variable aleatoria o distribución de probabilidad también se les denomina la media y la varianza de la población correspondiente.

En el campo de la inferencia estadística, el estadístico se interesa en llegar a conclusiones respecto a una población, cuando es imposible o poco práctico conocer todo el conjunto de observaciones que la constituyen. Por ejemplo, al intentar determinar la longitud de la vida promedio de cierta marca de bombilla, sería imposible probarlas todas si tenemos que dejar algunas para venderlas. Los costos desmesurados que implicaría estudiar a toda la población también constituirían un factor que impediría hacerlo. Por lo tanto, debemos depender de un subconjunto de observaciones de la población que nos ayude a realizar inferencias respecto a ella. Esto nos lleva a considerar el concepto de muestreo.

Definición 8.2: Una **muestra** es un subconjunto de una población.

Para que las inferencias que hacemos sobre la población a partir de la muestra sean válidas, debemos obtener muestras que sean representativas de ella. Con mucha

frecuencia nos sentimos tentados a elegir una muestra seleccionando a los miembros más convenientes de la población. Tal procedimiento podría conducir a inferencias erróneas respecto a la población. Se dice que cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población está **sesgado**. Para eliminar cualquier posibilidad de sesgo en el procedimiento de muestreo es deseable elegir una **muestra aleatoria**, lo cual significa que las observaciones se realicen de forma independiente y al azar.

Para seleccionar una muestra aleatoria de tamaño n de una población $f(x)$ definimos la variable aleatoria X_i, $i = 1, 2,..., n$, que representa la i-ésima medición o valor de la muestra que observamos. Si las mediciones se obtienen repitiendo el experimento n veces independientes en, esencialmente, las mismas condiciones, las variables aleatorias $X_1, X_2,..., X_n$ constituirán entonces una muestra aleatoria de la población $f(x)$ con valores numéricos $x_1, x_2,..., x_n$. Debido a las condiciones idénticas en las que se seleccionan los elementos de la muestra, es razonable suponer que las n variables aleatorias $X_1, X_2,..., X_n$ son independientes y que cada una tiene la misma distribución de probabilidad $f(x)$. Es decir, las distribuciones de probabilidad de $X_1, X_2,..., X_n$son, respectivamente, $f(x_1), f(x_2),..., f(x_n)$, y su distribución de probabilidad conjunta es $f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$. El concepto de muestra aleatoria se describe de manera formal en la siguiente definición.

Definición 8.3: Sean $X_1, X_2,..., X_n$ variables aleatorias independientes n, cada una con la misma distribución de probabilidad $f(x)$. Definimos $X_1, X_2,..., X_n$ como una **muestra aleatoria** de tamaño n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

Si se realiza una selección aleatoria de $n = 8$ baterías de almacenamiento de un proceso de fabricación que mantiene las mismas especificaciones, y al registrar la duración de cada batería se encuentra que la primera medición x_1 es un valor de X_1, la segunda medición x_2 es un valor de X_2, y así sucesivamente, entonces $x_1, x_2,..., x_8$ son los valores de la muestra aleatoria $X_1, X_2,..., X_8$. Si suponemos que la población de vidas útiles de las baterías es normal, los valores posibles de cualquier X_i, $i = 1, 2,..., 8$ serán exactamente los mismos que los de la población original, por consiguiente, X_i tiene una distribución normal idéntica a la de X.

8.2 Algunos estadísticos importantes

Nuestro principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros desconocidos de la población. Suponga, por ejemplo, que deseamos concluir algo respecto a la proporción de consumidores de café en Estados Unidos que prefieren cierta marca de café. Sería imposible interrogar a cada consumidor estadounidense de café para calcular el valor del parámetro p que representa la proporción de la población. En vez de esto se selecciona una muestra aleatoria grande y se calcula la proporción $\hat{p}$ de personas en esta muestra que prefieren la marca de café en cuestión. El valor $\hat{p}$ se utiliza ahora para hacer una inferencia respecto a la proporción p verdadera.

Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria; ya que es posible tomar muchas muestras aleatorias de la misma población, esperaríamos

que $\hat{p}$ variara un poco de una a otra muestra. Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P. Tal variable aleatoria se llama **estadístico**.

Definición 8.4: Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama **estadístico**.

Medidas de localización de una muestra: la media, la mediana y la moda muestrales

En el capítulo 4 presentamos los parámetros μ y σ^2, que miden el centro y la variabilidad de una distribución de probabilidad. Éstos son parámetros de población constantes y de ninguna manera se ven afectados o influidos por las observaciones de una muestra aleatoria. Definiremos, sin embargo, algunos estadísticos importantes que describen las medidas correspondientes de una muestra aleatoria. Los estadísticos que más se utilizan para medir el centro de un conjunto de datos, acomodados en orden de magnitud, son la **media**, la **mediana** y la **moda**. Aunque los primeros dos estadísticos se expusieron en el capítulo 1, repetiremos las definiciones. Sean $X_1, X_2, ..., X_n$ representaciones de n variables aleatorias.

a) Media muestral:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Observe que el estadístico $\bar{X}$ toma el valor $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ cuando X_1 toma el valor x_1, X_2 toma el valor x_2 y así sucesivamente. El término *media muestral* se aplica tanto al estadístico $\bar{X}$ como a su valor calculado $\bar{x}$.

b) Mediana muestral:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$

La mediana muestral también es una medida de localización que indica el valor central de la muestra. En la sección 1.3 se presentan ejemplos de la media muestral y de la mediana muestral. La moda muestral se define de la siguiente manera:

c) La moda muestral es el valor que ocurre con mayor frecuencia en la muestra.

Ejemplo 8.1: Suponga que un conjunto de datos consta de las siguientes observaciones:

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43

La moda de la muestra es 0.43, ya que este valor aparece con más frecuencia que los demás. ∎

Como se expuso en el capítulo 1, una medida de localización o tendencia central en una muestra no da por sí misma una indicación clara de la naturaleza de ésta, de manera que también debe considerarse una medida de variabilidad en la muestra.

Las medidas de variabilidad de una muestra: la varianza, la desviación estándar y el rango de la muestra

La variabilidad en la muestra refleja cómo se dispersan las observaciones a partir del promedio. Se remite al lector al capítulo 1 para un análisis más amplio. Es posible tener dos conjuntos de observaciones con las mismas media o mediana que difieran de manera considerable en la variabilidad de sus mediciones sobre el promedio.

Considere las siguientes mediciones, en litros, para dos muestras de jugo de naranja envasado por las empresas *A* y *B*:

Muestra *A*	0.97	1.00	0.94	1.03	1.06
Muestra *B*	1.06	1.01	0.88	0.91	1.14

Ambas muestras tienen la misma media, 1.00 litros. Es muy evidente que la empresa *A* envasa el jugo de naranja con un contenido más uniforme que la *B*. Decimos que la **variabilidad** o la **dispersión** de las observaciones a partir del promedio es menor para la muestra *A* que para la muestra *B*. Por lo tanto, al comprar jugo de naranja, tendríamos más confianza en que el envase que seleccionemos se acerque al promedio anunciado si se lo compramos a la empresa *A*.

En el capítulo 1 presentamos varias medidas de la variabilidad de una muestra, como la **varianza muestral**, la **desviación estándar muestral** y el **rango de la muestra**. En este capítulo nos enfocaremos sobre todo en la varianza de la muestra. Nuevamente, sea que $X_1, X_2, ..., X_n$ representan n variables aleatorias.

a) La varianza muestral:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{8.2.1}$$

El valor calculado de S^2 para una muestra dada se denota con s^2. Observe que S^2 se define esencialmente como el promedio de los cuadrados de las desviaciones de las observaciones a partir de su media. La razón para utilizar $n - 1$ como divisor, en vez de la elección más obvia n, quedará más clara en el capítulo 9.

Ejemplo 8.2: Una comparación de los precios de café en 4 tiendas de abarrotes de San Diego, seleccionadas al azar, mostró aumentos en comparación con el mes anterior de 12, 15, 17 y 20 centavos por bolsa de una libra. Calcule la varianza de esta muestra aleatoria de aumentos de precio.

***Solución*:** Si calculamos la media de la muestra, obtenemos

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ centavos.}$$

Por lo tanto,

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$
$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

Mientras que la expresión para la varianza de la muestra de la definición 8.6 ilustra mejor que S^2 es una medida de variabilidad, una expresión alternativa tiene cierto mérito, de manera que el lector debería conocerla. El siguiente teorema contiene tal expresión.

Teorema 8.1: Si S^2 es la varianza de una muestra aleatoria de tamaño n, podemos escribir

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

***Prueba*:** Por definición,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right].$$

Como en el capítulo 1, a continuación se definen la **desviación estándar muestral** y el **rango muestral**:

b) Desviación estándar muestral:

$$S = \sqrt{S^2},$$

donde S^2 es la varianza muestral.

Permitamos que $X_{\text{máx}}$ denote el más grande de los valores X_i y $X_{\text{mín}}$ el más pequeño.

c) Rango muestral:

$$R = X_{\text{máx}} - X_{\text{mín}}.$$

Ejemplo 8.3: Calcule la varianza de los datos 3, 4, 5, 6, 6 y 7, que representan el número de truchas atrapadas por una muestra aleatoria de 6 pescadores, el 19 de junio de 1996, en el lago Muskoka.

***Solución*:** Encontramos que $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$ y $n = 6$. De aquí,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Por consiguiente, la desviación estándar de la muestra $s = \sqrt{13/6} = 1.47$ y el rango muestral es $7 - 3 = 4$.

Ejercicios

8.1 Defina las poblaciones adecuadas a partir de las cuales se seleccionaron las siguientes muestras:

a) Se llamó por teléfono a personas de 200 casas en la ciudad de Richmond y se les pidió nombrar al candidato por el que votarían en la elección del presidente de la mesa directiva de la escuela.

b) Se lanzó 100 veces una moneda y se registraron 34 cruces.

c) Se probaron 200 pares de un nuevo tipo de calzado deportivo en un torneo de tenis profesional para determinar su duración y se encontró que, en promedio, duraron 4 meses.

d) En cinco ocasiones diferentes a una abogada le tomó 21, 26, 24, 22 y 21 minutos conducir desde su casa en los suburbios hasta su oficina en el centro de la ciudad.

8.2 El tiempo, en minutos, que 10 pacientes esperan en un consultorio médico antes de recibir tratamiento se registraron como sigue: 5, 11, 9, 5, 10, 15, 6, 10, 5 y 10. Trate los datos como una muestra aleatoria y calcule

a) la media;
b) la mediana;
c) la moda.

8.3 Los tiempos que los 9 individuos de una muestra aleatoria tardan en reaccionar ante un estimulante se registraron como 2.5, 3.6, 3.1, 4.3, 2.9, 2.3, 2.6, 4.1 y 3.4 segundos. Calcule

a) la media;
b) la mediana.

8.4 El número de multas emitidas por infracciones de tránsito por 8 oficiales estatales durante el fin de semana del día en Conmemoración de los Caídos es 5, 4, 7, 7, 6, 3, 8 y 6.

a) Si estos valores representan el número de multas emitidas por una muestra aleatoria de 8 oficiales estatales del condado de Montgomery, en Virginia, defina una población adecuada.
b) Si los valores representan el número de multas emitidas por una muestra aleatoria de 8 oficiales estatales de Carolina del Sur, defina una población adecuada.

8.5 El número de respuestas incorrectas en un examen de competencia de verdadero-falso para una muestra aleatoria de 15 estudiantes se registraron de la siguiente manera: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4 y 2. Calcule

a) la media;
b) la mediana;
c) la moda.

8.6 Calcule la media, la mediana y la moda para la muestra, cuyas observaciones, 15, 7, 8, 95, 19, 12, 8, 22 y 14 representan el número de días de incapacidad médica reportados en 9 solicitudes de devolución de impuestos. ¿Qué valor parece ser la mejor medida del centro de esos datos? Explique las razones de su preferencia.

8.7 Una muestra aleatoria de empleados de una fábrica local prometieron los siguientes donativos, en dólares, al United Fund: 100, 40, 75, 15, 20, 100, 75, 50, 30, 10, 55, 75, 25, 50, 90, 80, 15, 25, 45 y 100. Calcule

a) la media;
b) la moda.

8.8 De acuerdo con la escritora ecologista Jacqueline Killeen, los fosfatos que contienen los detergentes de uso casero pasan directamente a nuestros sistemas de desagüe, ocasionando que los lagos se conviertan en pantanos, los cuales a la larga se volverán desiertos. Los siguientes datos muestran la cantidad de fosfatos por carga de lavado, en gramos, para una muestra aleatoria de diversos tipos de detergentes que se usan de acuerdo con las instrucciones prescritas:

Detergente para ropa	Fosfatos por carga (gramos)
A & P Blue Sail	48
Dash	47
Concentrated All	42
Cold Water All	42
Breeze	41
Oxydol	34
Ajax	31
Sears	30
Fab	29
Cold Power	29
Bold	29
Rinso	26

Para los datos de fosfato dados, calcule

a) la media;
b) la mediana;
c) la moda.

8.9 Considere los datos del ejercicio 8.2 y calcule

a) el rango;
b) la desviación estándar.

8.10 Para la muestra de tiempos de reacción del ejercicio 8.3 calcule

a) el rango;
b) la varianza, utilizando la fórmula de la forma (8.2.1).

8.11 Para los datos del ejercicio 8.5 calcule la varianza utilizando la fórmula

a) de la forma (8.2.1);
b) del teorema 8.1.

8.12 El contenido de alquitrán de 8 marcas de cigarrillos que se seleccionan al azar de la lista más reciente publicada por la Comisión Federal de Comercio es el siguiente: 7.3, 8.6, 10.4, 16.1, 12.2, 15.1, 14.5 y 9.3 miligramos. Calcule

a) la media;
b) la varianza.

8.13 Los promedios de calificaciones de 20 estudiantes universitarios del último año, seleccionados al azar de una clase que se va a graduar, son los siguientes:

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

Calcule la desviación estándar.

8.14 *a*) Demuestre que la varianza de la muestra permanece sin cambio si a cada valor de la muestra se le suma o se le resta una constante c.

b) Demuestre que la varianza de la muestra se vuelve c^2 veces su valor original si cada observación de la muestra se multiplica por c.

8.15 Verifique que la varianza de la muestra 4, 9, 3, 6, 4 y 7 es 5.1, y utilice este hecho, junto con los resultados del ejercicio 8.14, para calcular

a) la varianza de la muestra 12, 27, 9, 18, 12 y 21;

b) la varianza de la muestra 9, 14, 8, 11, 9 y 12.

8.16 En la temporada 2004-2005 el equipo de futbol americano de la Universidad del Sur de California tuvo las siguientes diferencias de puntuación en los 13 partidos que jugó.

11 49 32 3 6 38 38 30 8 4 31 5 36

Calcule

a) la media de la diferencia de puntos;

b) la mediana de las diferencias de puntos.

8.3 Distribuciones muestrales

El campo de la inferencia estadística trata básicamente con generalizaciones y predicciones. Por ejemplo, con base en las opiniones de varias personas entrevistadas en la calle, los estadounidenses podrían afirmar que en una próxima elección 60% de los votantes de la ciudad de Detroit favorecerían a cierto candidato. En este caso tratamos con una muestra aleatoria de opiniones de una población finita muy grande. Por otro lado, con base en las estimaciones de 3 contratistas seleccionados al azar, de los 30 que laboran actualmente en esta ciudad, podríamos afirmar que el costo promedio de construir una residencia en Charleston, Carolina del Sur, está entre $330,000 y $335,000. La población que se va a muestrear aquí también es finita, pero muy pequeña. Finalmente, consideremos una máquina despachadora de bebida gaseosa que está diseñada para servir en promedio 240 mililitros de bebida. Un ejecutivo de la empresa calcula la media de 40 bebidas servidas y obtiene $\bar{x} = 236$ mililitros y, con base en este valor, decide que la máquina está sirviendo bebidas con un contenido promedio de $\mu = 240$ mililitros. Las 40 bebidas servidas representan una muestra de la población infinita de posibles bebidas que despachará esta máquina.

Inferencias sobre la población a partir de información de la muestra

En cada uno de los ejemplos anteriores calculamos un estadístico de una muestra que se selecciona de la población, y con base en tales estadísticos hicimos varias afirmaciones respecto a los valores de los parámetros de la población, que pueden ser o no ciertas. El ejecutivo de la empresa decide que la máquina despachadora está sirviendo bebidas con un contenido promedio de 240 mililitros, aunque la media de la muestra fue de 236 mililitros, porque conoce la teoría del muestreo según la cual, si $\mu = 240$ mililitros, tal valor de la muestra podría ocurrir fácilmente. De hecho, si realiza pruebas similares, cada hora por ejemplo, esperaría que los valores del estadístico $\bar{x}$ fluctuaran por arriba y por abajo de $\mu = 240$ mililitros. Sólo cuando el valor de $\bar{x}$ difiera considerablemente de 240 mililitros el ejecutivo de la empresa tomará medidas para ajustar la máquina.

Como un estadístico es una variable aleatoria que depende sólo de la muestra observada, debe tener una distribución de probabilidad.

Definición 8.5: La distribución de probabilidad de un estadístico se denomina **distribución muestral**.

La distribución muestral de un estadístico depende de la distribución de la población, del tamaño de las muestras y del método de selección de las muestras. En lo que resta de este capítulo estudiaremos varias de las distribuciones muestrales más importantes de los estadísticos que se utilizan con frecuencia. Las aplicaciones de tales distribuciones muestrales a problemas de inferencia estadística se consideran en la mayoría de los capítulos posteriores. La distribución de probabilidad de $\bar{X}$ se llama **distribución muestral de la media**.

¿Qué es la distribución muestral de $\bar{X}$?

Se deberían considerar las distribuciones muestrales de $\bar{X}$ y S^2 como los mecanismos a partir de los cuales se puede hacer inferencias acerca de los parámetros μ y σ^2. La distribución muestral de $\bar{X}$ con tamaño muestral n es la distribución que resulta cuando un **experimento se lleva a cabo una y otra vez** (siempre con una muestra de tamaño n) **y resultan los diversos valores de** $\bar{X}$. Por lo tanto, esta distribución muestral describe la variabilidad de los promedios muestrales alrededor de la media de la población μ. En el caso de la máquina despachadora de bebidas, el conocer la distribución muestral de $\bar{X}$ le permite al analista encontrar una discrepancia "típica" entre un valor $\bar{x}$ observado y el verdadero valor de μ. Se aplica el mismo principio en el caso de la distribución de S^2. La distribución muestral produce información acerca de la variabilidad de los valores de s^2 alrededor de σ^2 en experimentos que se repiten.

8.4 Distribución muestral de medias y el teorema del límite central

La primera distribución muestral importante a considerar es la de la media $\bar{X}$. Suponga que de una población normal con media μ y varianza σ^2 se toma una muestra aleatoria de n observaciones. Cada observación X_i, $i = 1, 2,..., n$, de la muestra aleatoria tendrá entonces la misma distribución normal que la población de donde se tomó. Así, por la propiedad reproductiva de la distribución normal que se estableció en el teorema 7.11, concluimos que

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

tiene una distribución normal con media

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ términos}}) = \mu \text{ y varianza } \sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ términos}}) = \frac{\sigma^2}{n}.$$

Si tomamos muestras de una población con distribución desconocida, ya sea finita o infinita, la distribución muestral de $\bar{X}$ aún será aproximadamente normal con media μ y varianza σ^2/n, siempre que el tamaño de la muestra sea grande. Este asombroso resultado es una consecuencia inmediata del siguiente teorema, que se conoce como teorema del límite central.

El teorema del límite central

Teorema 8.2: **Teorema del límite central:** Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n, tomada de una población con media μ y varianza finita σ^2, entonces la forma límite de la distribución de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

a medida que $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

La aproximación normal para $\bar{X}$ por lo general será buena si $n \geq 30$, siempre y cuando la distribución de la población no sea muy asimétrica. Si $n < 30$, la aproximación será buena sólo si la población no es muy diferente de una distribución normal y, como antes se estableció, si se sabe que la población es normal, la distribución muestral de $\bar{X}$ seguirá siendo una distribución normal exacta, sin importar qué tan pequeño sea el tamaño de las muestras.

El tamaño de la muestra $n = 30$ es un lineamiento para el teorema del límite central. Sin embargo, como indica el planteamiento del teorema, la suposición de normalidad en la distribución de $\bar{X}$ se vuelve más precisa a medida que n se hace más grande. De hecho, la figura 8.1 ilustra cómo funciona el teorema. La figura indica cómo la distribución de $\bar{X}$ se acerca más a la normalidad a medida que aumenta n, empezando con la distribución claramente asimétrica de una observación individual ($n = 1$). También ilustra que la media de $\bar{X}$ sigue siendo μ para cualquier tamaño de la muestra y que la varianza de $\bar{X}$ se vuelve más pequeña a medida que aumenta n.

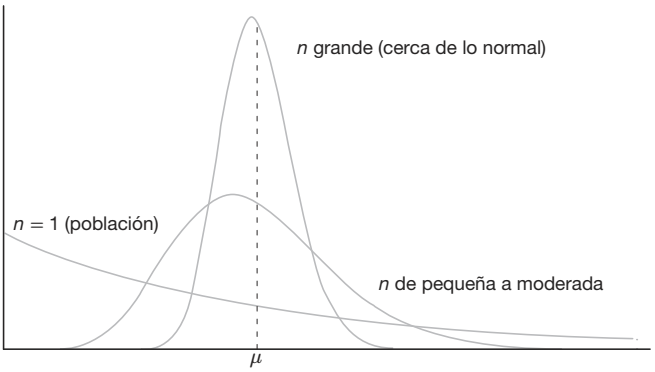


Figura 8.1: Ejemplo del teorema del límite central (distribución de $\bar{X}$ para $n = 1$, n moderada y n grande).

Ejemplo 8.4: Una empresa de material eléctrico fabrica bombillas que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Calcule la probabilidad de que una muestra aleatoria de 16 bombillas tenga una vida promedio de menos de 775 horas.

***Solución*:** La distribución muestral de $\bar{X}$ será aproximadamente normal, con $\mu_{\bar{X}} = 800$ y $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. La probabilidad que se desea es determinada por el área de la región sombreada de la figura 8.2.

En lo que corresponde a $\bar{x} = 775$, obtenemos que

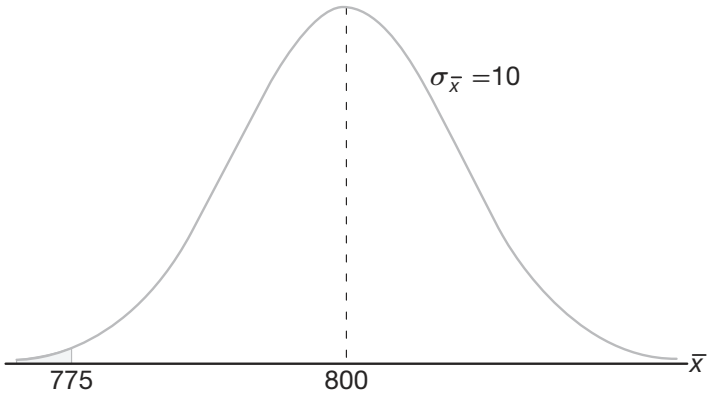


Figura 8.2: Área para el ejemplo 8.4.

$$z = \frac{775 - 800}{10} = -2.5,$$

y, por lo tanto,

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Inferencias sobre la media de la población

Una aplicación muy importante del teorema del límite central consiste en determinar valores razonables de la media de la población μ. Temas como prueba de hipótesis, estimación, control de calidad y muchos otros utilizan el teorema del límite central. El siguiente ejemplo ilustra cómo se utiliza el teorema del límite central con respecto a su relación con μ, la media poblacional, aunque la aplicación formal de los temas precedentes se deja para capítulos posteriores.

En el siguiente estudio de caso proporcionamos un ejemplo en el que se hace una inferencia utilizando la distribución muestral de $\bar{X}$. En este ejemplo sencillo se conocen μ y σ. El teorema del límite central y el concepto general de las distribuciones muestrales a menudo se utilizan para proporcionar evidencias acerca de algún aspecto importante de una distribución, por ejemplo uno de sus parámetros. En el caso del teorema del límite central el parámetro que nos interesa es la media μ. La inferencia que se hace acerca de μ puede adoptar una de varias formas. Con frecuencia el analista desea que los datos (en la forma de $\bar{x}$) respalden (o no) alguna conjetura predeterminada respecto al valor de μ. El uso de lo que sabemos sobre la distribución de muestreo puede contribuir a responder este tipo de pregunta. En el siguiente estudio de caso el concepto de prueba de hipótesis conduce a un objetivo formal que destacaremos en capítulos posteriores.

Estudio de caso 8.1: **Partes para automóviles.** Un importante proceso de fabricación produce partes de componentes cilíndricos para la industria automotriz. Es importante que el proceso produzca partes que tengan un diámetro medio de 5.0 milímetros. El ingeniero implicado asume

que la media de la población es de 5.0 milímetros. Se lleva a cabo un experimento donde se seleccionan al azar 100 partes elaboradas por el proceso y se mide el diámetro de cada una de ellas. Se sabe que la desviación estándar de la población es $\sigma = 0.1$ milímetros. El experimento indica un diámetro promedio muestral de $\bar{x} = 5.027$ milímetros. ¿Esta información de la muestra parece apoyar o refutar la suposición del ingeniero?

***Solución*:** Este ejemplo refleja el tipo de problemas que a menudo se presentan y que se resuelven con las herramientas de pruebas de hipótesis que se presentan en los siguientes capítulos. No utilizaremos aquí el formalismo asociado con la prueba de hipótesis, pero ilustraremos los principios y la lógica que se utilizan.

El hecho de que los datos apoyen o refuten la suposición depende de la probabilidad de que datos similares a los que se obtuvieron en este experimento ($\bar{x} = 5.027$) pueden ocurrir con facilidad cuando de hecho $\mu = 5.0$ (figura 8.3). En otras palabras, ¿qué tan probable es que se pueda obtener $\bar{x} \geq 5.027$ con $n = 100$, si la media de la población es $\mu = 5.0$? Si esta probabilidad sugiere que $\bar{x} = 5.027$ no es poco razonable, no se refuta la suposición. Si la probabilidad es muy baja, se puede argumentar con certidumbre que los datos no apoyan la suposición de que $\mu = 5.0$. La probabilidad que elegimos para el cálculo es dada por $P(|\bar{X} - 5| \geq 0.027)$.

En otras palabras, si la media μ es 5, ¿cuál es la probabilidad de que $\bar{X}$ se desvíe

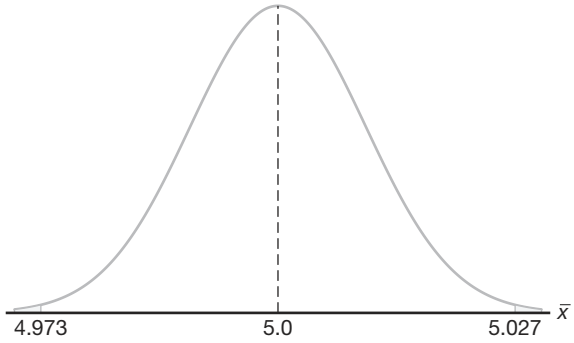


Figura 8.3: Área para el estudio de caso 8.1.

cuando mucho hasta 0.027 milímetros?

$$P(|\bar{X} - 5| \geq 0.027) = P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq -0.027)$$
$$= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right).$$

Aquí simplemente estandarizamos $\bar{X}$ de acuerdo con el teorema del límite central. Si la suposición $\mu = 5.0$ es cierta, $\frac{\bar{X} - 5}{0.1/\sqrt{100}}$ debería ser $N(0, 1)$. Por consiguiente,

$$2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right) = 2P(Z \geq 2.7) = 2(0.0035) = 0.007.$$

Por lo tanto, se experimentaría por casualidad que una $\bar{x}$ estaría a 0.027 milímetros

de la media en tan sólo 7 de 1000 experimentos. Como resultado, este experimento con $\bar{x} = 5.027$ ciertamente no ofrece evidencia que apoye la suposición de que $\mu = 5.0$. De hecho, ¡la refuta consistentemente!

Ejemplo 8.5: El viaje en un autobús especial para ir de un campus de una universidad al campus de otra en una ciudad toma, en promedio, 28 minutos, con una desviación estándar de 5 minutos. En cierta semana un autobús hizo el viaje 40 veces. ¿Cuál es la probabilidad de que el tiempo promedio del viaje sea mayor a 30 minutos? Suponga que el tiempo promedio se redondea al entero más cercano.

***Solución*:** En este caso $\mu = 28$ y $\sigma = 3$. Necesitamos calcular la probabilidad $P(\bar{X} > 30)$ con $n = 40$. Como el tiempo se mide en una escala continua redondeada al minuto más cercano, una $\bar{x}$ mayor que 30 sería equivalente a $\bar{x} \geq 30.5$. Por lo tanto,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

Hay sólo una ligera probabilidad de que el tiempo promedio de un viaje del autobús exceda 30 minutos. En la figura 8.4 se presenta una gráfica ilustrativa.

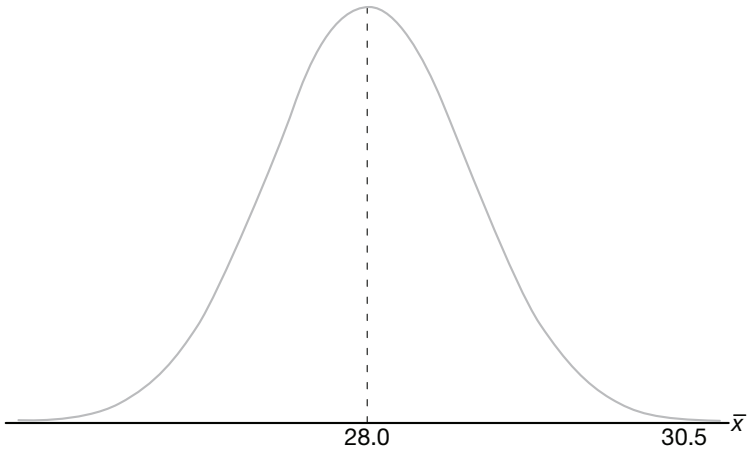


Figura 8.4: Área para el ejemplo 8.5.

Distribución muestral de la diferencia entre dos medias

La ilustración del estudio de caso 8.1 se refiere a conceptos de inferencia estadística sobre una sola media μ. El ingeniero estaba interesado en respaldar una suposición con respecto a una sola media de población. Una aplicación mucho más importante incluye dos poblaciones. Un científico o ingeniero se podrían interesar en un experimento donde se comparan dos métodos de producción: el 1 y el 2. La base para tal comparación es $\mu_1 - \mu_2$, la diferencia entre las medias de población.

Suponga que tenemos dos poblaciones, la primera con media μ_1 y varianza σ_1^2, y la segunda con media μ_2 y varianza σ_2^2. Representemos con el estadístico $\bar{X}_1$ la media

de una muestra aleatoria de tamaño n_1, seleccionada de la primera población, y con el estadístico $\bar{X}_2$ la media de una muestra aleatoria de tamaño n_2 seleccionada de la segunda población, independiente de la muestra de la primera población. ¿Qué podríamos decir acerca de la distribución muestral de la diferencia $\bar{X}_1 - \bar{X}_2$ para muestras repetidas de tamaños n_1 y n_2? De acuerdo con el teorema 8.2, tanto la variable $\bar{X}_1$ como la variable $\bar{X}_2$ están distribuidas más o menos de forma normal con medias μ_1 y μ_2 y varianzas σ_1^2/n_1 y σ_2^2/n_2, respectivamente. Esta aproximación mejora a medida que aumentan n_1 y n_2. Al elegir muestras independientes de las dos poblaciones nos aseguramos de que las variables $\bar{X}_1$ y $\bar{X}_2$ sean independientes y, usando el teorema 7.11, con $a_1 = 1$ y $a_2 = -1$, concluimos que $\bar{X}_1 - \bar{X}_2$ se distribuye aproximadamente de forma normal con media

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

y varianza

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

El teorema del límite central se puede ampliar fácilmente al caso de dos muestras y dos poblaciones.

Teorema 8.3: Si se extraen al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente, entonces la distribución muestral de las diferencias de las medias, $\bar{X}_1 - \bar{X}_2$, tiene una distribución aproximadamente normal, con media y varianza dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ y } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De aquí,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

es aproximadamente una variable normal estándar.

Si tanto n_1 como n_2 son mayores o iguales que 30, la aproximación normal para la distribución de $\bar{X}_1 - \bar{X}_2$ es muy buena cuando las distribuciones subyacentes no están tan alejadas de la normal. Sin embargo, aun cuando n_1 y n_2 sean menores que 30, la aproximación normal es hasta cierto punto buena, excepto cuando las poblaciones no son definitivamente normales. Por supuesto, si ambas poblaciones son normales, entonces $\bar{X}_1 - \bar{X}_2$ tiene una distribución normal sin importar de qué tamaño sean n_1 y n_2.

La utilidad de la distribución muestral de la diferencia entre los dos promedios muestrales es muy similar a la que se describe en el estudio de caso 8.1 en la página 235 para el caso de una sola media. Ahora presentaremos el estudio de caso 8.2, que se enfoca en el uso de la diferencia entre dos medias muestrales para respaldar (o no) la suposición de que dos medias de población son iguales.

Estudio de caso 8.2: **Tiempo de secado de pinturas.** Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura, el *A* y el *B*. Con la pintura tipo *A* se pintan 18 especímenes y se registra el tiempo (en horas) que cada uno tarda en secar. Lo mismo se hace con la pintura tipo *B*. Se sabe que la desviación estándar de población de ambas es 1.0.

Si se supone que los especímenes pintados se secan en el mismo tiempo medio con los dos tipos de pintura, calcule $P(\bar{X}_A - \bar{X}_B > 1.0)$, donde $\bar{X}_A$ y $\bar{X}_B$ son los tiempos promedio de secado para muestras de tamaño $n_A = n_B = 18$.

***Solución*:** A partir de la distribución de muestreo de $\bar{X}_A - \bar{X}_B$ sabemos que la distribución es aproximadamente normal con media

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$$

y varianza

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

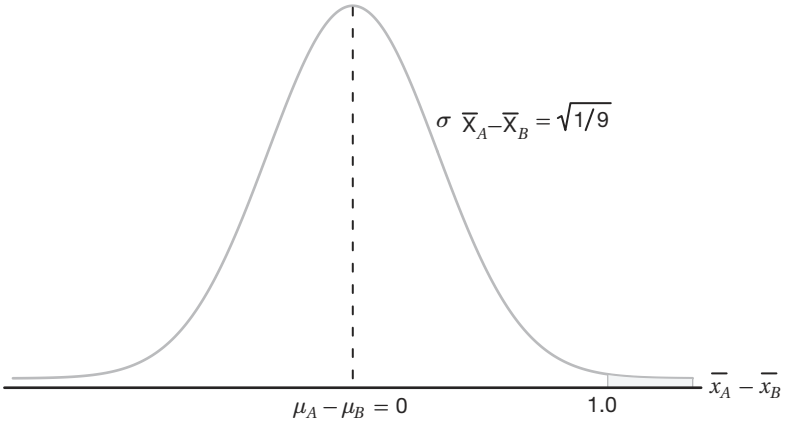


Figura 8.5: Área para el estudio de caso 8.2.

La probabilidad que se desea es dada por la región sombreada en la figura 8.5. En correspondencia con el valor $\bar{X}_A - \bar{X}_B = 1.0$, tenemos

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

de modo que

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

¿Qué aprendemos del estudio de caso 8.2?

La mecánica en el cálculo se basa en la suposición de que $\mu_A = \mu_B$. Suponga, sin embargo, que el experimento realmente se lleva a cabo con el fin de hacer una inferencia respecto a la igualdad de μ_A y μ_B, los tiempos medios de secado de las dos poblaciones. Si se encontrara que los dos promedios difieren por una hora (o más), este resultado sería una evidencia que nos llevaría a concluir que el tiempo medio de secado de la población

no es igual para los dos tipos de pintura. Por otro lado, suponga que la diferencia en los dos promedios muestrales es tan pequeña como, digamos, 15 minutos. Si $\mu_A = \mu_B$,

$$P[(\bar{X}_A - \bar{X}_B) > 0.25 \text{ horas}] = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{\sqrt{1/9}} > \frac{3}{4}\right)$$
$$= P\left(Z > \frac{3}{4}\right) = 1 - P(Z < 0.75) = 1 - 0.7734 = 0.2266.$$

Como esta probabilidad no es baja, se concluiría que una diferencia de 15 minutos en las medias de las muestras puede ocurrir por azar, es decir, sucede con frecuencia aunque $\mu_A = \mu_B$. Por lo tanto, este tipo de diferencia en el tiempo promedio de secado ciertamente *no es una señal clara* de que $\mu_A \neq \mu_B$.

Como indicamos al principio, en los capítulos siguientes se observará un formalismo más detallado con respecto a éste y a otros tipos de inferencia estadística, por ejemplo, la prueba de hipótesis. El teorema del límite central y las distribuciones de muestreo que se presentan en las siguientes tres secciones también desempeñarán un papel fundamental.

Ejemplo 8.6: Los cinescopios para televisor del fabricante *A* tienen una duración media de 6.5 años y una desviación estándar de 0.9 años; mientras que los del fabricante *B* tienen una duración media de 6.0 años y una desviación estándar de 0.8 años. ¿Cuál es la probabilidad de que una muestra aleatoria de 36 cinescopios del fabricante *A* tenga por lo menos 1 año más de vida media que una muestra de 49 cinescopios del fabricante *B*?

***Solución*:** Tenemos la siguiente información:

Población 1	Población 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

Si utilizamos el teorema 8.3, la distribución muestral de $\bar{X}_1 - \bar{X}_2$ será aproximadamente normal y tendrá una media y una desviación estándar de

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{y} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

La probabilidad de que 36 cinescopios del fabricante *A* tengan por lo menos 1 año más de vida media que 49 cinescopios del fabricante *B* es dada por el área de la región sombreada de la figura 8.6. Con respecto al valor $\bar{x}_1 - \bar{x}_2 = 1.0$, encontramos que

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

y de aquí

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65)$$
$$= 1 - 0.9960 = 0.0040.$$

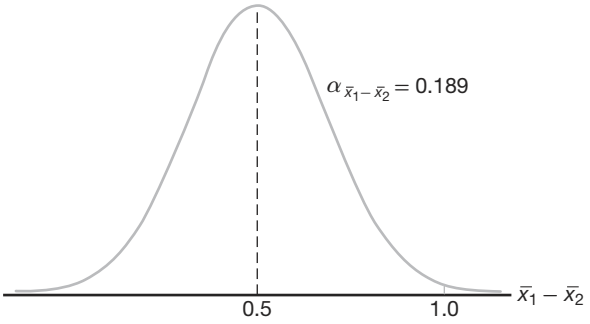


Figura 8.6: Área para el ejemplo 8.6.

Más sobre la distribución muestral de medias. Aproximación normal a la distribución binomial

En la sección 6.5 analizamos a fondo la aproximación normal a la distribución binomial. Estaban dadas las condiciones sobre los parámetros n y p, para los cuales la distribución de una variable aleatoria binomial se puede aproximar mediante la distribución normal. Los ejemplos y los ejercicios reflejaron la importancia del concepto de "aproximación normal". Resulta que el teorema del límite central da más idea de cómo y por qué funciona esta aproximación. Sabemos con certeza que una variable aleatoria binomial es el número X de éxitos en n pruebas independientes, donde el resultado de cada prueba es binario. En el capítulo 1 también vimos que la proporción calculada en un experimento así es un promedio de un conjunto de ceros y unos. De hecho, mientras que la proporción X/n es un promedio, X es la suma de este conjunto de ceros y unos, y tanto X como X/n son casi normales si n es suficientemente grande. Desde luego, a partir de lo que aprendimos en el capítulo 6, sabemos que hay condiciones de n y p que afectan la calidad de la aproximación; a saber, $np \geq 5$ y $nq \geq 5$.

Ejercicios

8.17 Si se extraen todas las muestras posibles de tamaño 16 de una población normal con media igual a 50 y desviación estándar igual a 5, ¿cuál es la probabilidad de que una media muestral $\bar{X}$ caiga en el intervalo que va de $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ a $\mu_{\bar{X}} - 0.4\sigma_{\bar{X}}$? Suponga que las medias muestrales se pueden medir con cualquier grado de precisión.

8.18 Si la desviación estándar de la media para la distribución muestral de muestras aleatorias de tamaño 36 de una población grande o infinita es 2, ¿qué tan grande debe ser el tamaño de la muestra si la desviación estándar se reduce a 1.2?

8.19 Se fabrica cierto tipo de hilo con una resistencia a la tensión media de 78.3 kilogramos y una desviación estándar de 5.6 kilogramos. ¿Cómo cambia la varianza de la media muestral cuando el tamaño de la muestra

a) aumenta de 64 a 196?
b) disminuye de 784 a 49?

8.20 Dada la población uniforme discreta

$$f(x) = \begin{cases} \frac{1}{3}, & x = 2, 4, 6, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule la probabilidad de que una muestra aleatoria de tamaño 54, seleccionada con reemplazo, produzca una media muestral mayor que 4.1 pero menor que 4.4. Suponga que las medias se miden al décimo más cercano.

8.21 Una máquina de bebidas gaseosas se ajusta de manera que la cantidad de bebida que sirve promedie 240 mililitros con una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 40 bebidas y calculando el

contenido promedio. Si la media de las 40 bebidas es un valor dentro del intervalo $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, se piensa que la máquina opera satisfactoriamente; de lo contrario, se ajusta. En la sección 8.3 el ejecutivo de la empresa encontró que la media de 40 bebidas era $\bar{x} = 236$ mililitros y concluyó que la máquina no necesitaba un ajuste. ¿Fue ésta una decisión razonable?

8.22 Las estaturas de 1000 estudiantes se distribuyen aproximadamente de forma normal con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se extraen 200 muestras aleatorias de tamaño 25 de esta población y las medias se registran al décimo de centímetro más cercano, determine

a) la media y la desviación estándar de la distribución muestral de $\bar{X}$;
b) el número de las medias muestrales que caen entre 172.5 y 175.8 centímetros;
c) el número de medias muestrales que caen por debajo de 172.0 centímetros.

8.23 La variable aleatoria X, que representa el número de cerezas en un tarta, tiene la siguiente distribución de probabilidad:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

a) Calcule la media μ y la varianza σ^2 de X.
b) Calcule la media $\mu_{\bar{X}}$ y la varianza $\sigma^2_{\bar{X}}$ de la media $\bar{X}$ para muestras aleatorias de 36 tartas de cereza.
c) Calcule la probabilidad de que el número promedio de cerezas en 36 tartas sea menor que 5.5.

8.24 Si cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms, ¿cuál es la probabilidad de que una muestra aleatoria de 36 de estas resistencias tenga una resistencia combinada de más de 1458 ohms?

8.25 La vida media de una máquina para elaborar pan es de 7 años, con una desviación estándar de 1 año. Suponga que la vida de estas máquinas sigue aproximadamente una distribución normal y calcule

a) la probabilidad de que la vida media de una muestra aleatoria de 9 de estas máquinas caiga entre 6.4 y 7.2 años;
b) el valor de x a la derecha del cual caería 15% de las medias calculadas de muestras aleatorias de tamaño 9.

8.26 La cantidad de tiempo que le toma al cajero de un banco con servicio en el automóvil atender a un cliente es una variable aleatoria con una media $\mu = 3.2$ minutos y una desviación estándar $\sigma = 1.6$ minutos. Si se observa una muestra aleatoria de 64 clientes, calcule la probabilidad de que el tiempo medio que el cliente pasa en la ventanilla del cajero sea

a) a lo sumo 2.7 minutos;
b) más de 3.5 minutos;
c) al menos 3.2 minutos pero menos de 3.4 minutos.

8.27 En un proceso químico la cantidad de cierto tipo de impureza en el producto es difícil de controlar y por ello es una variable aleatoria. Se especula que la cantidad media de la población de impurezas es 0.20 gramos por gramo del producto. Se sabe que la desviación estándar es 0.1 gramos por gramo. Se realiza un experimento para entender mejor la especulación de que $\mu = 0.2$. El proceso se lleva a cabo 50 veces en un laboratorio y el promedio de la muestra $\bar{x}$ resulta ser 0.23 gramos por gramo. Comente sobre la especulación de que la cantidad media de impurezas es 0.20 gramos por gramo. Utilice el teorema del límite central en su respuesta.

8.28 Se toma una muestra aleatoria de tamaño 25 de una población normal que tiene una media de 80 y una desviación estándar de 5. Una segunda muestra aleatoria de tamaño 36 se toma de una población normal diferente que tiene una media de 75 y una desviación estándar de 3. Calcule la probabilidad de que la media muestral calculada de las 25 mediciones exceda la media muestral calculada de las 36 mediciones por lo menos 3.4 pero menos de 5.9. Suponga que las diferencias de las medias se miden al décimo más cercano.

8.29 La distribución de alturas de cierta raza de perros *terrier* tiene una media de 72 centímetros y una desviación estándar de 10 centímetros; en tanto que la distribución de alturas de cierta raza de *poodles* tiene una media de 28 centímetros con una desviación estándar de 5 centímetros. Suponga que las medias muestrales se pueden medir con cualquier grado de precisión y calcule la probabilidad de que la media muestral de una muestra aleatoria de alturas de 64 *terriers* exceda la media muestral para una muestra aleatoria de alturas de 100 *poodles* a lo sumo 44.2 centímetros.

8.30 La calificación promedio de los estudiantes de primer año en un examen de aptitudes en cierta universidad es 540, con una desviación estándar de 50. Suponga que las medias se miden con cualquier grado de precisión. ¿Cuál es la probabilidad de que dos grupos seleccionados al azar, que constan de 32 y 50 estudiantes, respectivamente, difieran en sus calificaciones promedio por

a) más de 20 puntos?
b) una cantidad entre 5 y 10 puntos?

8.31 Considere el estudio de caso 8.2 de la página 238. Suponga que en un experimento se utilizaron 18 especímenes para cada tipo de pintura y que $\bar{x}_A - \bar{x}_B$, la diferencia real en el tiempo medio de secado, resultó ser 1.0.

a) ¿Parecería ser un resultado razonable si los dos tiempos promedio de secado de las dos poblaciones realmente son iguales? Utilice el resultado que se obtuvo en el estudio de caso 8.2.

b) Si alguien hiciera el experimento 10,000 veces bajo la condición de que $\mu_A = \mu_B$, ¿en cuántos de esos 10,000 experimentos habría una diferencia $\bar{x}_A - \bar{x}_B$ tan grande como 1.0 (o más grande)?

8.32 Dos máquinas diferentes de llenado de cajas se utilizan para llenar cajas de cereal en una línea de ensamble. La medición fundamental en la que influyen estas máquinas es el peso del producto en las cajas. Los ingenieros están seguros de que la varianza en el peso del producto es $\sigma^2 = 1$ onza. Se realizan experimentos usando ambas máquinas con tamaños muestrales de 36 cada una. Los promedios muestrales para las máquinas A y B son $\bar{x}_A = 4.5$ onzas y $\bar{x}_B = 4.7$ onzas. Los ingenieros se sorprenden de que los dos promedios maestrales para las máquinas de llenado sean tan diferentes.

a) Utilice el teorema del límite central para determinar

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

bajo la condición de que $\mu_A = \mu_B$.

b) ¿Los experimentos mencionados parecen, de cualquier forma, apoyar consistentemente la suposición de que las medias de población de las dos máquinas son diferentes? Explique utilizando la respuesta que encontró en el inciso *a*.

8.33 El benceno es una sustancia química altamente tóxica para los seres humanos. Sin embargo, se utiliza en la fabricación de medicamentos, de tintes y de recubrimientos, así como en la peletería. Las regulaciones del gobierno establecen que el contenido de benceno en el agua que resulte de cualquier proceso de producción en el que participe esta sustancia no debe exceder 7950 partes por millón (ppm). Para un proceso particular de interés, un fabricante recolectó una muestra de agua 25 veces de manera aleatoria y el promedio muestral $\bar{x}$ fue de 7960 ppm. A partir de los datos históricos, se sabe que la desviación estándar σ es 100 ppm.

a) ¿Cuál es la probabilidad de que el promedio muestral en este experimento exceda el límite establecido por el gobierno, si la media de la población es igual al límite? Utilice el teorema del límite central.

b) ¿La $\bar{x} = 7960$ observada en este experimento es firme evidencia de que la media de la población en este proceso excede el límite impuesto por el gobierno? Responda calculando

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Suponga que la distribución de la concentración de benceno es normal.

8.34 En la fabricación de cierto producto de acero se están utilizando dos aleaciones, la *A* y la *B*. Se necesita diseñar un experimento para comparar las dos aleaciones en términos de su capacidad de carga máxima en toneladas, es decir, la cantidad máxima de carga que pueden soportar sin romperse. Se sabe que las dos desviaciones estándar de la capacidad de carga son iguales a 5 toneladas cada una. Se realiza un experimento en el que se prueban 30 especímenes de cada aleación (*A* y *B*) y se obtienen los siguientes resultados:

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

Los fabricantes de la aleación *A* están convencidos de que esta evidencia demuestra de forma concluyente que $\mu_A > \mu_B$ y, por lo tanto, que su aleación es mejor. Los fabricantes de la aleación *B* afirman que el experimento fácilmente podría haber resultado $\bar{x}_A - \bar{x}_B = 4$, *incluso si* las dos medias de población fueran iguales. En otras palabras, "¡los resultados no son concluyentes!".

a) Encuentre un argumento que ponga en evidencia el error de los fabricantes de la aleación *B*. Para ello calcule

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

b) ¿Considera que estos datos apoyan fuertemente a la aleación *A*?

8.35 Considere la situación del ejemplo 8.4 de la página 234. ¿Los resultados que se obtuvieron allí lo llevan a cuestionar la premisa de que $\mu = 800$ horas? Proporcione un resultado probabilístico que indique qué tan *raro* es el evento $\bar{X} \leq 775$ cuando $\mu = 800$. Por otro lado, ¿qué tan raro sería si μ fuera, verdaderamente, digamos, $\neq 760$ horas?

8.36 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución que sólo puede adoptar valores positivos. Utilice el teorema del límite central para argumentar que si n es tan grande como se requiere, entonces $Y = X_1 X_2 \ldots X_n$ tiene aproximadamente una distribución logarítmica normal.

8.5 Distribución muestral de S^2

En la sección anterior aprendimos acerca de la distribución muestral de $\bar{X}$. El teorema del límite central nos permitió utilizar el hecho de que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiende a $N(0, 1)$ a medida que crece el tamaño de la muestra. *Las distribuciones muestrales de estadísticos importantes* nos permiten conocer información sobre los parámetros. Por lo general, los parámetros son las contrapartes del estadístico en cuestión. Por ejemplo, si un ingeniero se interesa en la resistencia media de la población de cierto tipo de resistencia, sacará provecho de la distribución muestral de $\bar{X}$ una vez que reúna la información de la muestra. Por otro lado, si está estudiando la variabilidad en la resistencia, evidentemente utilizará la distribución muestral de S^2 para conocer la contraparte paramétrica, la varianza de la población σ^2.

Si se extrae una muestra aleatoria de tamaño n de una población normal con media μ y varianza σ^2, y se calcula la varianza muestral, se obtiene un valor del estadístico S^2. Procederemos a considerar la distribución del estadístico $(n-1)S^2/\sigma^2$.

Mediante la suma y la resta de la media muestral $\bar{X}$ es fácil ver que

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Al dividir cada término de la igualdad entre σ^2 y sustituir $(n-1)S^2$ por $\sum_{i=1}^{n}(X_i-\bar{X})^2$, obtenemos

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Ahora, de acuerdo con el corolario 7.1 de la página 222, sabemos que

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

es una variable aleatoria chi cuadrada con n grados de libertad. Tenemos una variable aleatoria chi cuadrada con n grados de libertad dividida en dos componentes. Observe que en la sección 6.7 demostramos que una distribución chi cuadrada es un caso especial de la distribución gamma. El segundo término del lado derecho es Z^2, que es una variable aleatoria chi cuadrada con 1 grado de libertad, y resulta que $(n-1)S^2/\sigma^2$ es una variable aleatoria chi cuadrada con $n-1$ grados de libertad. Formalizamos esto en el siguiente teorema.

Teorema 8.4: Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces el estadístico

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v=n-1$ grados de libertad.

Los valores de la variable aleatoria χ^2 se calculan de cada muestra mediante la fórmula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

La probabilidad de que una muestra aleatoria produzca un valor χ^2 mayor que algún valor específico es igual al área bajo la curva a la derecha de este valor. El valor χ^2 por arriba del cual se encuentra un área de α por lo general se representa con χ^2_α. Esto se ilustra mediante la región sombreada de la figura 8.7.

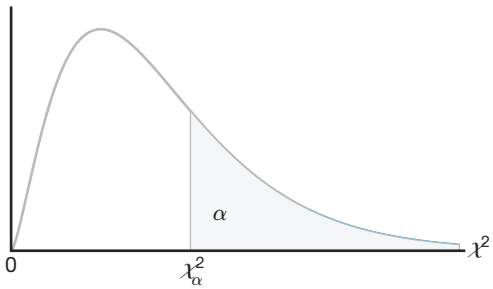


Figura 8.7: La distribución chi cuadrada.

La tabla A.5 da los valores de χ^2_α para diversos valores de α y v. Las áreas, α, son los encabezados de las columnas; los grados de libertad, v, se dan en la columna izquierda, y las entradas de la tabla son los valores χ^2. En consecuencia, el valor χ^2 con 7 grados de libertad, que deja un área de 0.05 a la derecha, es $\chi^2_{0.05} = 14.067$. Debido a la falta de simetría, para encontrar $\chi^2_{0.95} = 2.167$ para $v = 7$ también debemos usar las tablas.

Exactamente 95% de una distribución chi cuadrada cae entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Un valor χ^2 que cae a la derecha de $\chi^2_{0.025}$ no tiene probabilidades de ocurrir, a menos que el valor de σ^2 que supusimos sea demasiado pequeño. Lo mismo sucede con un valor χ^2 que cae a la izquierda de $\chi^2_{0.975}$, el cual tampoco es probable que ocurra, a menos que el valor de σ^2 que supusimos sea demasiado grande. En otras palabras, es posible tener un valor χ^2 a la izquierda de $\chi^2_{0.975}$ o a la derecha de $\chi^2_{0.025}$ cuando el valor de σ^2 es correcto; pero si esto sucediera, lo más probable es que el valor de σ^2 que se supuso sea un error.

Ejemplo 8.7: Un fabricante de baterías para automóvil garantiza que su producto durará, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, ¿el fabricante continuará convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que las duraciones de las baterías siguen una distribución normal.

***Solución*:** Primero se calcula la varianza de la muestra usando el teorema 8.1,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Entonces,

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

es un valor de una distribución chi cuadrada con 4 grados de libertad. Como 95% de los valores χ^2 con 4 grados de libertad cae entre 0.484 y 11.143, el valor calculado con $\sigma^2 = 1$ es razonable y, por lo tanto, el fabricante no tiene razones para sospechar que la desviación estándar no sea igual a 1 año. ■

Grados de libertad como una medición de la información muestral

Del corolario 7.1 expuesto en la sección 7.3 recuerde que

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

tiene una distribución χ^2 con *n grados de libertad*. Observe también el teorema 8.4, el cual indica que la variable aleatoria

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución χ^2 con $n - 1$ *grados de libertad*. El lector debe también recordar que el término *grados de libertad*, que se utiliza en este contexto idéntico, se estudió en el capítulo 1.

Como antes indicamos, el teorema 8.4 no se demostrará; sin embargo, el lector puede verlo como una indicación de que cuando no se conoce μ y se considera la distribución de

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

hay **1 grado menos de libertad**, o se pierde un grado de libertad al estimar μ (es decir, cuando μ se reemplaza por $\bar{x}$). En otras palabras, en la muestra aleatoria de la distribución normal hay n grados de libertad o *partes de información* independientes. Cuando los datos (los valores en la muestra) se utilizan para calcular la media, hay un grado menos de libertad en la información que se utiliza para estimar σ^2.

8.6 Distribución *t*

En la sección 8.4 se analizó la utilidad del teorema del límite central. Sus aplicaciones giran en torno a las inferencias sobre una media de la población o a la diferencia entre dos medias de población. En este contexto es evidente la utilidad de utilizar el teorema del límite central y la distribución normal. Sin embargo, se supuso que se conoce la desviación estándar de la población. Esta suposición quizá sea razonable en situaciones en las que el ingeniero está muy familiarizado con el sistema o proceso. Sin embargo, en muchos escenarios experimentales el conocimiento de σ no es ciertamente más razonable que el conocimiento de la media de la población μ. A menudo, de hecho, una estimación de σ debe ser proporcionada por la misma información muestral que produce el promedio muestral $\bar{x}$. Como resultado, un estadístico natural a considerar para tratar con las inferencias sobre μ es

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

dado que S es el análogo de la muestra para σ. Si el tamaño de la muestra es pequeño, los valores de S^2 fluctúan de forma considerable de una muestra a otra (véase el ejercicio 8.43 de la página 259) y la distribución de T se desvía de forma apreciable de la de una distribución normal estándar.

Si el tamaño de la muestra es suficientemente grande, digamos $n \geq 30$, la distribución de T no difiere mucho de la normal estándar. Sin embargo, para $n < 30$ es útil tratar con la distribución exacta de T. Para desarrollar la distribución muestral de T, supondremos que nuestra muestra aleatoria se seleccionó de una población normal. Podemos escribir, entonces,

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}},$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiene una distribución normal estándar y

$$V = \frac{(n-1)S^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n - 1$ grados de libertad. Al obtener muestras de poblaciones normales se puede demostrar que $\bar{X}$ y S^2 son independientes y, en consecuencia, también lo son Z y V. El siguiente teorema proporciona la definición de una variable aleatoria T como una función de Z (normal estándar) y χ^2. Para completar se proporciona la función de densidad de la distribución t.

Teorema 8.5: Sea Z una variable aleatoria normal estándar y V una variable aleatoria chi cuadrada con v grados de libertad. Si Z y V son independientes, entonces la distribución de la variable aleatoria T, donde

$$T = \frac{Z}{\sqrt{V/v}},$$

es dada por la función de densidad

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

Ésta se conoce como la **distribución *t*** con v grados de libertad.

A partir de lo antes expuesto, y del teorema anterior, se deriva el siguiente corolario.

Corolario 8.1: Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes normales con media μ y desviación estándar σ. Sea

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Entonces la variable aleatoria $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ tiene una distribución t con $v = n - 1$ grados de libertad.

La distribución de probabilidad de T se publicó por primera vez en 1908 en un artículo de W. S. Gosset. En esa época, Gosset trabajaba para una cervecería irlandesa que prohibía a sus empleados que publicaran los resultados de sus investigaciones. Para evadir la prohibición Gosset publicó su trabajo en secreto bajo el seudónimo de "Student". Es por esto que a la distribución de T se le suele llamar distribución t de Student o simplemente distribución t. Para derivar la ecuación de esta distribución Gosset supuso que las muestras se seleccionaban de una población normal. Aunque ésta parecería una suposición muy restrictiva, se puede demostrar que las poblaciones que no son normales y que poseen distribuciones en forma casi de campana aún proporcionan valores de T que se aproximan muy de cerca a la distribución t.

¿Qué apariencia tiene la distribución *t*?

La distribución de T se parece a la distribución de Z en que ambas son simétricas alrededor de una media de cero. Ambas distribuciones tienen forma de campana, pero la distribución t es más variable debido al hecho de que los valores T dependen de las fluctuaciones de dos cantidades, $\bar{X}$ y S^2; mientras que los valores Z dependen sólo de los cambios en $\bar{X}$ de una muestra a otra. La distribución de T difiere de la de Z en que la varianza de T depende del tamaño de la muestra n y siempre es mayor que 1. Sólo cuando el tamaño de la muestra $n \to \infty$ las dos distribuciones serán iguales. En la figura 8.8 se presenta la relación entre una distribución normal estándar ($v = \infty$) y las distribuciones t con 2 y 5 grados de libertad. Los puntos porcentuales de la distribución t se dan en la tabla A.4.

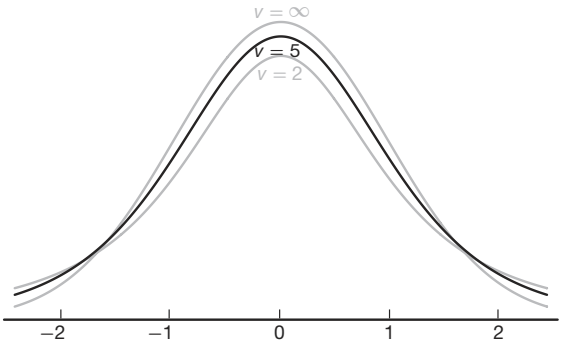


Figura 8.8: Curvas de la distribución t para $v = 2, 5$ y ∞.

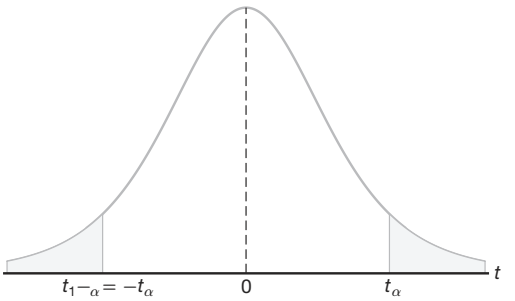


Figura 8.9: Propiedad de simetría (alrededor de 0) de la distribución t.

El valor t por arriba del cual se encuentra un área igual a α por lo general se representa con t_α. Por consiguiente, el valor t con 10 grados de libertad que deja una área de 0.025 a la derecha es $t = 2.228$. Como la distribución t es simétrica alrededor de una media de cero, tenemos $t_{1-\alpha} = -t_\alpha$; es decir, el valor t que deja una área de $1 - \alpha$ a la derecha y, por lo tanto, una área de α a la izquierda es igual al valor t negativo que deja una área de α en la cola derecha de la distribución (véase la figura 8.9). Esto es, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, etcétera.

Ejemplo 8.8: El valor t con $v = 14$ grados de libertad que deja una área de 0.025 a la izquierda y, por lo tanto, una área de 0.975 a la derecha, es

$$t_{0.975} = -t_{0.025} = -2.145.$$

Ejemplo 8.9: Calcule $P(-t_{0.025} < T < t_{0.05})$.

***Solución*:** Como $t_{0.05}$ deja una área de 0.05 a la derecha y $-t_{0.025}$ deja una área de 0.025 a la izquierda, obtenemos una área total de

$$1 - 0.05 - 0.025 = 0.925$$

entre $-t_{0.025}$ y $t_{0.05}$. En consecuencia,

$$P(-t_{0.025} < T < t_{0.05}) = 0.925.$$

Ejemplo 8.10: Calcule k tal que $P(k < T < -1.761) = 0.045$ para una muestra aleatoria de tamaño 15 que se selecciona de una distribución normal y $\frac{\overline{X} - \mu}{s/\sqrt{n}}$.

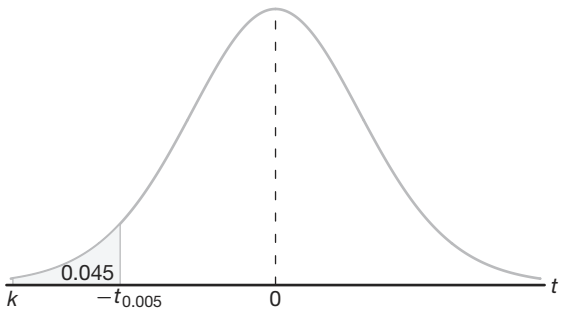


Figura 8.10: Valores t para el ejemplo 8.10.

***Solución*:** A partir de la tabla A.4 advertimos que 1.761 corresponde a $t_{0.05}$ cuando $v = 14$. Por lo tanto, $-t_{0.05} = -1.761$. Puesto que en el enunciado de probabilidad original k está a la izquierda de $-t_{0.05} = -1.761$, tenemos que $k = -t_\alpha$. Entonces, a partir de la figura 8.10, tenemos

$$0.045 = 0.05 - \alpha, \text{ o } \alpha = 0.005.$$

Así, de la tabla A.4 con $v = 14$,

$$k = -t_{0.005} = -2.977 \text{ y } P(-2.977 < T < -1.761) = 0.045.$$

Exactamente 95% de los valores de una distribución t con $v = n - 1$ grados de libertad caen entre $-t_{0.025}$ y $t_{0.025}$. Por supuesto, hay otros valores t que contienen 95% de la distribución, como $-t_{0.02}$ y $t_{0.03}$, pero estos valores no aparecen en la tabla A.4 y, además, el intervalo más corto posible se obtiene eligiendo valores t que dejen exactamente la misma área en las dos colas de nuestra distribución. Un valor t que caiga por debajo de $-t_{0.025}$ o por arriba de $t_{0.025}$ tendería a hacernos creer que ha ocurrido un evento muy raro, o que quizá nuestra suposición acerca de μ es un error. Si esto ocurriera, tendríamos que tomar la decisión de que el valor de μ que supusimos es erróneo. De hecho, un valor t que cae por debajo de $-t_{0.01}$ o por arriba de $t_{0.01}$ proporcionaría incluso evidencia más sólida de que el valor de μ que supusimos es muy improbable. En el capítulo 10 se tratarán procedimientos generales para probar aseveraciones respecto al valor del parámetro μ. El siguiente ejemplo ilustra una vista preliminar del fundamento de tales procedimientos.

Ejemplo 8.11: Un ingeniero químico afirma que el rendimiento medio de la población de un cierto proceso de lotes es 500 gramos por mililitro de materia prima. Para verificar dicha afirmación muestrea 25 lotes cada mes. Si el valor t calculado cae entre $-t_{0.05}$ y $t_{0.05}$, queda satisfecho con su afirmación. ¿Qué conclusión debería sacar de una muestra que tiene una media $\bar{x} = 518$ gramos por mililitro y una desviación estándar muestral $s = 40$ gramos? Suponga que la distribución de rendimientos es aproximadamente normal.

***Solución*:** En la tabla A.4 encontramos que $t_{0.05} = 1.711$ para 24 grados de libertad. Por lo tanto, el ingeniero quedará satisfecho con esta afirmación si una muestra de 25 lotes rinde un valor t entre −1.711 y 1.711. Si $\mu = 500$, entonces,

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

un valor muy superior a 1.711. La probabilidad de obtener un valor t, con $v = 24$, igual o mayor que 2.25, es aproximadamente 0.02. Si $\mu > 500$, el valor de t calculado de la muestra sería más razonable. Por lo tanto, es probable que el ingeniero concluya que el proceso produce un mejor producto del que pensaba. ∎

¿Para qué se utiliza la distribución *t*?

La distribución t se usa ampliamente en problemas relacionados con inferencias acerca de la media de la población (como se ilustra en el ejemplo 8.11) o en problemas que implican muestras comparativas (es decir, en casos donde se trata de determinar si las medias de dos muestras son muy diferentes). El uso de la distribución se ampliará en los capítulos 9, 10, 11 y 12. El lector debería notar que el uso de la distribución t para el estadístico

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

requiere que $X_1, X_2, \ldots, X_n$ sean normales. El uso de la distribución t y la consideración del tamaño de la muestra no se relacionan con el teorema del límite central. El uso de la distribución normal estándar en vez de T para $n \geq 30$ sólo implica, en este caso, que S es un estimador suficientemente bueno de σ. En los siguientes capítulos la distribución t se usa con amplitud.

8.7 Distribución *F*

Recomendamos la distribución *t* en parte por su aplicación a problemas en los que hay muestreo comparativo, es decir, a problemas en que se tienen que comparar dos medias muestrales. Por ejemplo, algunos de los ejemplos que daremos en los siguientes capítulos adoptarán un método aún más formal; un ingeniero químico reúne datos de dos catalizadores, un biólogo recoge datos sobre dos medios de crecimiento o un químico reúne datos sobre dos métodos de recubrimiento de material para prevenir la corrosión. Si bien es importante que la información muestral aclare lo relacionado con dos medias de población, a menudo éste es el caso en el que comparar la variabilidad es igual de importante, si no es que más. La distribución *F* tiene una amplia aplicación en la comparación de varianzas muestrales y también es aplicable en problemas que implican dos o más muestras.

El estadístico *F* se define como el cociente de dos variables aleatorias chi cuadrada independientes, dividida cada una entre su número de grados de libertad. En consecuencia, podemos escribir

$$F = \frac{U/v_1}{V/v_2},$$

donde U y V son variables aleatorias independientes que tienen distribuciones chi cuadrada con v_1 y v_2 grados de libertad, respectivamente. Estableceremos ahora la distribución muestral de *F*.

Teorema 8.6: Sean U y V dos variables aleatorias independientes que tienen distribuciones chi cuadrada con v_1 y v_2 grados de libertad, respectivamente. Entonces, la distribución de la variable aleatoria $F = \frac{U/v_1}{V/v_2}$ es dada por la función de densidad

$$h(f) = \begin{cases} \dfrac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \dfrac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

Ésta se conoce como la **distribución *F*** con v_1 y v_2 grados de libertad (g.l.).

En capítulos posteriores utilizaremos ampliamente la variable aleatoria *F*. Sin embargo, no emplearemos la función de densidad, la cual sólo se dará como complemento. La curva de la distribución *F* no sólo depende de los dos parámetros v_1 y v_2 sino también del orden en el que se establecen. Una vez que tenemos estos dos valores, podemos identificar la curva. En la figura 8.11 se presentan distribuciones *F* típicas.

Sea f_α el valor f por arriba del cual encontramos un área igual a α. Esto se ilustra mediante la región sombreada de la figura 8.12. La tabla A.6 proporciona valores de f_α sólo para $\alpha = 0.05$ y $\alpha = 0.01$ para varias combinaciones de los grados de libertad v_1 y v_2. Por lo tanto, el valor f con 6 y 10 grados de libertad, que deja un área de 0.05 a la derecha, es $f_{0.05} = 3.22$. Por medio del siguiente teorema, la tabla A.6 también se puede utilizar para encontrar valores de $f_{0.95}$ y $f_{0.99}$. La demostración se deja al lector.

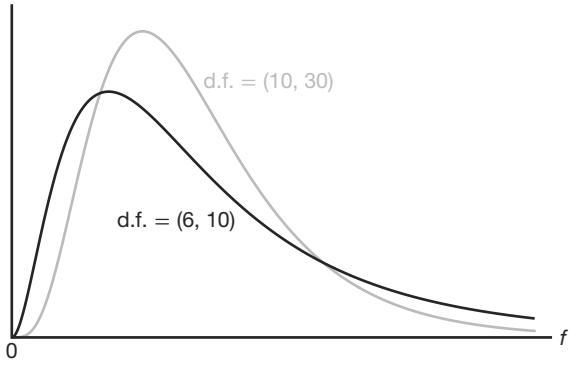


Figura 8.11: Distribuciones F típicas.

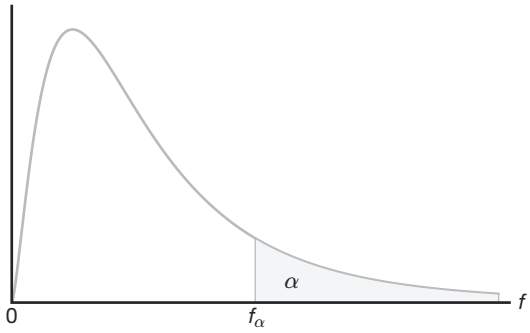


Figura 8.12: Ilustración de la f_α para la distribución F.

Teorema 8.7: Al escribir $f_\alpha(v_1, v_2)$ para f_α con v_1 y v_2 grados de libertad, obtenemos

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

Por consiguiente, el valor f con 6 y 10 grados de libertad, que deja una área de 0.95 a la derecha, es

$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246.$$

La distribución F con dos varianzas muestrales

Suponga que las muestras aleatorias de tamaños n_1 y n_2 se seleccionan de dos poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente. Del teorema 8.4, sabemos que

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ y } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

son variables aleatorias que tienen distribuciones chi cuadrada con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Además, como las muestras se seleccionan al azar, tratamos con variables aleatorias independientes. Entonces, usando el teorema 8.6 con $\chi_1^2 = U$ y $\chi_2^2 = V$, obtenemos el siguiente resultado.

Teorema 8.8: Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente, entonces,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

¿Para qué se utiliza la distribución *F*?

Al inicio de esta sección contestamos esta pregunta parcialmente. La distribución *F* se usa en situaciones de dos muestras para hacer inferencias acerca de las varianzas de población, lo cual implica aplicar el teorema 8.8. Sin embargo, la distribución *F* también se puede aplicar a muchos otros tipos de problemas que involucren varianzas muestrales. De hecho, la distribución *F* se llama *distribución de razón de varianzas*. Como ejemplo, considere el estudio de caso 8.2 en el que se compararon las dos pinturas, *A* y *B*, en relación con el tiempo medio que tardan en secar, en donde la distribución normal se aplica muy bien (suponiendo que se conocen σ_A y σ_B). Sin embargo, suponga que necesitamos comparar tres tipos de pinturas, digamos *A*, *B* y *C,* y que queremos determinar si las medias de población son equivalentes. Suponga que un resumen de la información importante del experimento es el siguiente:

Pintura	Media muestral	Varianza muestral	Tamaño muestral
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.20$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

El problema se centra alrededor de si los promedios muestrales ($\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$) están o no suficientemente alejados. La implicación de "suficientemente alejados" resulta muy importante. Parecería razonable que si la variabilidad entre los promedios muestrales es mayor que lo que se esperaría por casualidad, los datos no apoyan la conclusión de que $\mu_A = \mu_B = \mu_C$. Si estos promedios muestrales pudieran ocurrir por casualidad depende de la *variabilidad dentro de las muestras*, cuando se cuantifican por medio de s_A^2, s_B^2 y s_C^2. La idea de los componentes importantes de la variabilidad se observa mejor utilizando algunas gráficas sencillas. Considere la gráfica de los datos brutos de las muestras *A*, *B* y *C* que se presenta en la figura 8.13. Estos datos podrían generar con facilidad la información antes resumida.

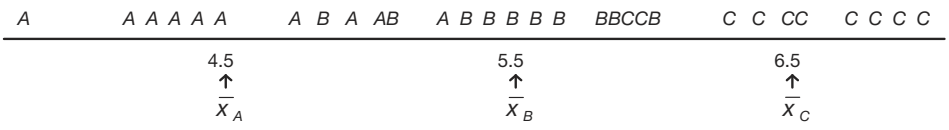


Figura 8.13: Datos de tres muestras diferentes.

Parece evidente que los datos provienen de distribuciones con diferentes medias de población, aunque hay cierto traslape entre las muestras. Un análisis que incluya todos los datos intentaría determinar si la variabilidad entre los promedios muestrales *y* la variabilidad dentro de las muestras podría haber ocurrido conjuntamente *si, de hecho, las poblaciones tienen una media común*. Observe que la clave para este análisis se centra alrededor de las dos siguientes fuentes de variabilidad.

1. Variabilidad dentro de las muestras (entre observaciones en muestras distintas).

2. Variabilidad entre muestras (entre promedios muestrales).

Es evidente que si la variabilidad en 1) es considerablemente mayor que en 2), entonces habrá un traslape considerable en los datos muestrales, una señal de que los datos podrían provenir de una distribución común. En el conjunto de datos que se presenta en la

figura 8.14 se encuentra un ejemplo. Por otro lado, es muy improbable que los datos de una distribución con una media común puedan tener una variabilidad entre promedios muestrales que sea considerablemente mayor que la variabilidad dentro de las muestras.

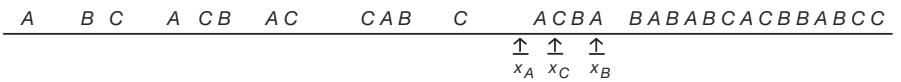


Figura 8.14: Datos que con facilidad podrían provenir de la misma población.

Las fuentes de variabilidad en 1) y 2) generan importantes cocientes de *varianzas muestrales* y los cocientes se utilizan junto con la distribución F. El procedimiento general implicado se llama **análisis de varianza**. Es interesante que en el ejemplo de la pintura aquí descrito tratamos con inferencias sobre tres medias de población pero utilizamos dos fuentes de variabilidad. No proporcionaremos detalles aquí, pero en los capítulos 13, 14 y 15 utilizaremos ampliamente el análisis de varianza en donde, por supuesto, la distribución F desempeña un papel importante.

8.8 Gráficas de cuantiles y de probabilidad

En el capítulo 1 presentamos al lector las distribuciones empíricas. El objetivo es utilizar presentaciones creativas para extraer información acerca de las propiedades de un conjunto de datos. Por ejemplo, los diagramas de tallo y hojas brindan al observador una imagen de la simetría y de otras propiedades de los datos. En este capítulo tratamos con muestras que, por supuesto, son conjuntos de datos experimentales de los que sacamos conclusiones sobre las poblaciones. A menudo, la apariencia de la muestra proporciona información sobre la distribución de la que se tomaron los datos. Por ejemplo, en el capítulo 1 ilustramos la naturaleza general de pares de muestras con gráficas de puntos que presentan una comparación relativa entre la tendencia central y la variabilidad de dos muestras.

En los capítulos siguientes con frecuencia supondremos que una distribución es normal. La información gráfica respecto a la validez de esta suposición se puede obtener a partir de presentaciones como los diagramas de tallo y hojas y los histogramas de frecuencias. Además, en esta sección presentaremos los conceptos de *gráficas de probabilidad normal* y *gráficas de cuantiles*. Estas gráficas se utilizan en estudios con diversos grados de complejidad con el principal objetivo de que las gráficas proporcionen una verificación diagnóstica sobre la suposición de que los datos provienen de una distribución normal.

Podemos caracterizar el análisis estadístico como el proceso de sacar conclusiones acerca de los sistemas en presencia de la variabilidad del sistema. Por ejemplo, el intento de un ingeniero por aprender acerca de un proceso químico a menudo es obstaculizado por la *variabilidad del proceso*. Un estudio que implica el número de artículos defectuosos en un proceso de producción con frecuencia se dificulta por la variabilidad en el método con el que se fabrican. En las secciones anteriores aprendimos acerca de las muestras y los estadísticos que expresan el centro de localización y la variabilidad en la muestra. Tales estadísticos ofrecen medidas simples, en tanto que una presentación gráfica brinda información adicional por medio de una imagen.

Un tipo de gráfica que puede ser especialmente útil para revelar la naturaleza de un conjunto de datos es la *gráfica de cuantiles*. Igual que en el caso de la gráfica de caja y extensión (véase la sección 1.6), en el que el objetivo del analista es hacer distinciones, en la gráfica de cuantiles se pueden utilizar las ideas básicas para *comparar muestras de*

datos. En los siguientes capítulos se presentarán más ejemplos del uso de las gráficas de cuantiles, en los que se analizará la inferencia estadística formal asociada con la comparación de muestras. En su momento, los estudios de caso mostrarán al lector tanto la inferencia formal como las gráficas diagnósticas para el mismo conjunto de datos.

Gráfica de cuantiles

El propósito de las gráficas de cuantiles consiste en describir, en forma de muestra, la función de distribución acumulada que se estudió en el capítulo 3.

Definición 8.6: Un **cuantil** de una muestra, $q(f)$, es un valor para el que una fracción específica f de los valores de los datos es menor que o igual a $q(f)$.

Evidentemente, un cuantil representa una estimación de una característica de una población o, más bien, la distribución teórica. La mediana de la muestra es $q(0.5)$. El percentil 75 (cuartil superior) es $q(0.75)$ y el cuartil inferior es $q(0.25)$.

Una **gráfica de cuantiles** simplemente *grafica los valores de los datos en el eje vertical contra una evaluación empírica de la fracción de observaciones excedidas por los valores de los datos*. Para propósitos teóricos esta fracción se calcula con

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}},$$

donde i es el orden de las observaciones cuando se ordenan de la menor a la mayor. En otras palabras, si denotamos las observaciones ordenadas como

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)},$$

entonces la gráfica de cuantiles describe una gráfica de $y_{(i)}$ contra f_i. En la figura 8.15 se presenta la gráfica de cuantiles para las asas de las latas de pintura analizadas con anterioridad.

A diferencia de la gráfica de caja y extensión, la gráfica de cuantiles realmente muestra todas las observaciones. Todos los cuantiles, incluidos la mediana y los cuantiles superior e inferior, se pueden aproximar de forma visual. Por ejemplo, observamos fácilmente una mediana de 35 y un cuartil superior de alrededor de 36. Las agrupaciones relativamente grandes en torno a valores específicos se indican por pendientes cercanas a cero; mientras que los datos escasos en ciertas áreas producen pendientes más abruptas. La figura 8.15 describe la dispersión de datos de los valores 28 a 30, pero una densidad relativamente alta de 36 a 38. En los capítulos 9 y 10 proseguimos con las gráficas de cuantiles mediante la ilustración de formas útiles en que es posible comparar distintas muestras.

Debería ser muy evidente para el lector que detectar si un conjunto de datos proviene o no de una distribución normal puede ser una herramienta importante para el analista de datos. Como antes indicamos en esta sección, a menudo suponemos que la totalidad o subconjuntos de las observaciones en un conjunto de datos son realizaciones de variables aleatorias normales independientes idénticamente distribuidas. Una vez más, la gráfica de diagnóstico a menudo se agrega a (con fines de presentación) una *prueba de bondad del ajuste* formal de los datos. Las pruebas de bondad del ajuste se estudiarán en el capítulo 10. Los lectores de un artículo o informe científico suelen considerar la información de diagnóstico mucho más clara, menos árida y quizá menos aburrida que un análisis formal.

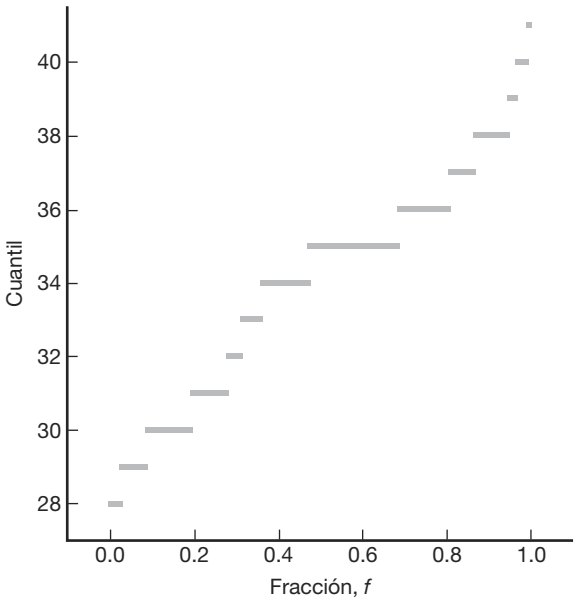


Figura 8.15: Gráfica de cuantiles para los datos de la pintura.

En los capítulos siguientes (del 9 al 13) nos enfocaremos nuevamente en los métodos de detección de desviaciones de la normalidad como un agregado de la inferencia estadística formal. Las gráficas de cuantiles son útiles para detectar los tipos de distribución. En la elaboración de modelos y en el diseño de experimentos también hay situaciones en que se utilizan las gráficas para detectar **términos** o **efectos del modelo** que están activos. En otras situaciones se utilizan para determinar si las suposiciones subyacentes que el científico o el ingeniero hicieron en la construcción del modelo son o no razonables. En los capítulos 11, 12 y 13 se incluyen muchos ejemplos con ilustraciones. La siguiente subsección brinda un análisis y un ejemplo de una gráfica de diagnóstico denominada *gráfica de cuantiles-cuantiles normales*.

Gráfica de cuantiles-cuantiles normales

La gráfica de cuantiles-cuantiles normales aprovecha lo que se conoce sobre los cuantiles de la distribución normal. La metodología incluye una gráfica de los cuantiles empíricos recién analizados, contra el cuantil correspondiente de la distribución normal. Ahora, la expresión para un cuantil de una variable aleatoria $N(\mu, \sigma)$ es muy complicada. Sin embargo, una buena aproximación es dada por

$$q_{\mu,\sigma}(f) = \mu + \sigma\{4.91[f^{0.14} - (1-f)^{0.14}]\}.$$

La expresión entre las llaves (el múltiplo de σ) es la aproximación para el cuantil correspondiente para la variable aleatoria $N(0, 1)$, es decir,

$$q_{0,1}(f) = 4.91[f^{0.14} - (1-f)^{0.14}].$$

Definición 8.7: La **gráfica de cuantiles-cuantiles normales** es una gráfica de $y_{(i)}$ (observaciones ordenadas) contra $q_{0,1}(f_i)$, donde $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$.

Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal. La intersección en el eje vertical es una estimación de la media de la población μ y la pendiente es una estimación de la desviación estándar σ. La figura 8.16 presenta una gráfica de cuantiles-cuantiles normales para los datos de las latas de pintura.

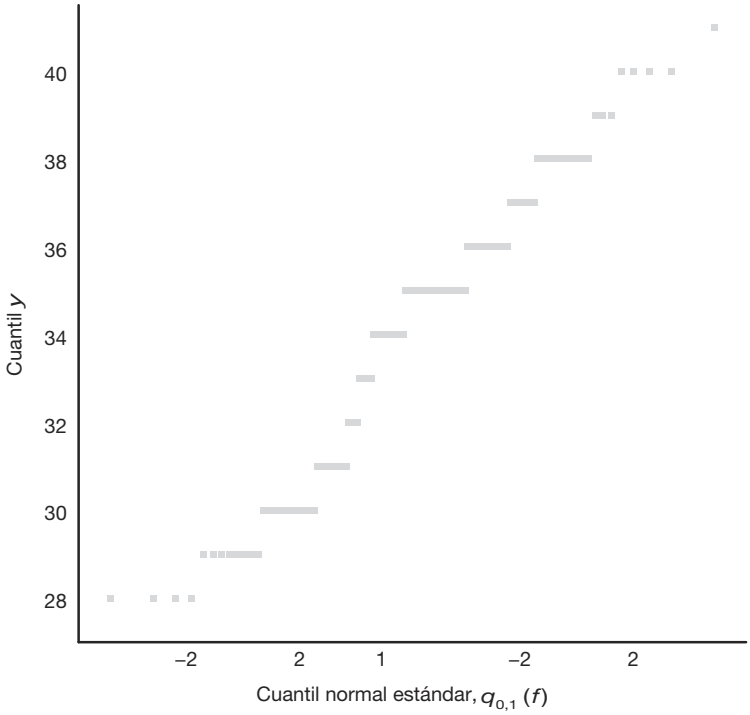


Figura 8.16: Gráfica de cuantiles-cuantiles normales para los datos de la pintura.

Graficación de la probabilidad normal

Observe cómo la desviación de la normalidad se vuelve evidente gracias a la apariencia de la gráfica. La asimetría que exhiben los datos produce cambios en la pendiente.

Las ideas para graficar la probabilidad se manifiestan en versiones diferentes de la gráfica de cuantiles-cuantiles normales que se presentó aquí. Por ejemplo, se ha puesto mucha atención a la llamada **gráfica de probabilidad normal**, en la que f se grafica contra los valores de los datos ordenados en un papel especial y la escala utilizada da como resultado una línea recta. Además, una gráfica alternativa utiliza los valores esperados de las observaciones clasificadas para la distribución normal y dibuja las observaciones clasificadas contra su valor esperado, bajo el supuesto de datos de $N(\mu, \sigma)$. Una vez más, la línea recta es el criterio gráfico que se emplea. Continuamos sugiriendo que basarse en los métodos analíticos gráficos que se describen en esta sección ayudará a comprender los métodos formales que permiten distinguir muestras diferentes de datos.

Ejemplo 8.12: Considere los datos del ejercicio 10.41 en la página 358 del capítulo 10. En el estudio "Retención de nutrientes y respuesta de comunidades de macroinvertebrados ante la presión de aguas residuales en un ecosistema fluvial", que se llevó a cabo en el departamento de zoología del Virginia Polytechnic Institute y la universidad estatal, se recabaron datos sobre mediciones de densidad (número de organismos por metro cuadrado) en dos diferentes estaciones colectoras. En el capítulo 10 se dan detalles con respecto a los métodos analíticos de comparación de muestras para determinar si ambas provienen de la misma distribución $N(\mu, \sigma)$. Los datos se presentan en la tabla 8.1.

Tabla 8.1: Datos para el ejemplo 8.12

Número de organismos por metro cuadrado			
Estación 1		**Estación 2**	
5,030	4,980	2,800	2,810
13,700	11,910	4,670	1,330
10,730	8,130	6,890	3,320
11,400	26,850	7,720	1,230
860	17,660	7,030	2,130
2,200	22,800	7,330	2,190
4,250	1,130		
15,040	1,690		

Dibuje una gráfica de cuantiles-cuantiles normales y saque conclusiones con respecto a si es razonable o no suponer que las dos muestras provienen de la misma distribución $n(x; \mu, \sigma)$.

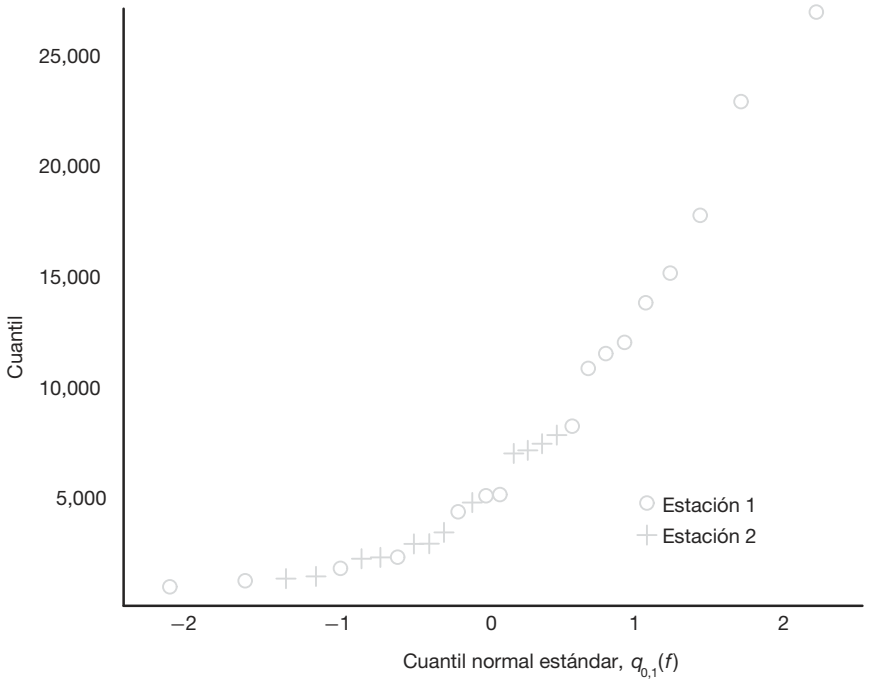


Figura 8.17: Gráfica de cuantiles-cuantiles normales para los datos de densidad del ejemplo 8.12.

***Solución*:** La figura 8.17 muestra la gráfica de cuantiles-cuantiles normales para las mediciones de densidad. La gráfica se aleja mucho de una sola línea recta. De hecho, los datos de la estación 1 reflejan pocos valores en la cola inferior de la distribución y varios en la cola superior. El "agrupamiento" de observaciones hace que parezca improbable que las dos muestras provengan de una distribución común $N(\mu, \sigma)$. ∎

Aunque hemos concentrado nuestra explicación y ejemplo en las gráficas de probabilidad para distribuciones normales, podemos enfocarnos en cualquier distribución. Tan sólo necesitaríamos calcular cantidades de forma analítica para la distribución teórica en cuestión.

Ejercicios

8.37 Para una distribución chi cuadrada calcule

a) $\chi^2_{0.025}$ cuando $v = 15$;

b) $\chi^2_{0.01}$ cuando $v = 7$;

c) $\chi^2_{0.05}$ cuando $v = 24$.

8.38 Para una distribución chi cuadrada, calcule

a) $\chi^2_{0.005}$ cuando $v = 5$;

b) $\chi^2_{0.05}$ cuando $v = 19$;

c) $\chi^2_{0.01}$ cuando $v = 12$.

8.39 Para una distribución chi cuadrada calcule χ^2_α, tal que

a) $P(X^2 > \chi^2_\alpha) = 0.99$ cuando $v = 4$;

b) $P(X^2 > \chi^2_\alpha) = 0.025$ cuando $v = 19$;

c) $P(37.652 < X^2 < \chi^2_\alpha) = 0.045$ cuando $v = 25$.

8.40 Para una distribución chi cuadrada calcule χ^2_α, tal que

a) $P(X^2 > \chi^2_\alpha) = 0.01$ cuando $v = 21$;

b) $P(X^2 < \chi^2_\alpha) = 0.95$ cuando $v = 6$;

c) $P(\chi^2_\alpha < X^2 < 23.209) = 0.015$ cuando $v = 10$.

8.41 Suponga que las varianzas muestrales son mediciones continuas. Calcule la probabilidad de que una muestra aleatoria de 25 observaciones, de una población normal con varianza $\sigma^2 = 6$, tenga una varianza muestral S^2

a) mayor que 9.1;

b) entre 3.462 y 10.745.

8.42 Las calificaciones de un examen de colocación que se aplicó a estudiantes de primer año de una universidad durante los últimos cinco años tienen una distribución aproximadamente normal con una media $\mu = 74$ y una varianza $\sigma^2 = 8$. ¿Seguiría considerando que $\sigma^2 = 8$ es un valor válido de la varianza si una muestra aleatoria de 20 estudiantes, a los que se les aplica el examen de colocación este año, obtienen un valor de $s^2 = 20$?

8.43 Demuestre que la varianza de S^2 para muestras aleatorias de tamaño n de una población normal disminuye a medida que aumenta n. [*Sugerencia*: primero calcule la varianza de $(n-1)S^2/\sigma^2$].

8.44 *a*) Calcule $t_{0.025}$ cuando $v = 14$.

b) Calcule $-t_{0.10}$ cuando $v = 10$.

c) Calcule $t_{0.995}$ cuando $v = 7$.

8.45 *a*) Calcule $P(T < 2.365)$ cuando $v = 7$.

b) Calcule $P(T > 1.318)$ cuando $v = 24$.

c) Calcule $P(-1.356 < T < 2.179)$ cuando $v = 12$.

d) Calcule $P(T > -2.567)$ cuando $v = 17$.

8.46 *a*) Calcule $P(-t_{0.005} < T < t_{0.01})$ para $v = 20$.

b) Calcule $P(T > -t_{0.025})$.

8.47 Dada una muestra aleatoria de tamaño 24 de una distribución normal, calcule k tal que

a) $P(-2.069 < T < k) = 0.965$;

b) $P(k < T < 2.807) = 0.095$;

c) $P(-k < T < k) = 0.90$.

8.48 Una empresa que fabrica juguetes electrónicos afirma que las baterías que utiliza en sus productos duran un promedio de 30 horas. Para mantener este promedio se prueban 16 baterías cada mes. Si el valor t calculado cae entre $-t_{0.025}$ y $t_{0.025}$, la empresa queda satisfecha con su afirmación. ¿Qué conclusiones debería sacar la empresa a partir de una muestra que tiene una media de $\bar{x} = 27.5$ horas y una desviación estándar de $s = 5$ horas? Suponga que la distribución de las duraciones de las baterías es aproximadamente normal.

8.49 Una población normal con varianza desconocida tiene una media de 20. ¿Es posible obtener una muestra aleatoria de tamaño 9 de esta población con una media de 24 y una desviación estándar de 4.1? Si no fuera posible, ¿a qué conclusión llegaría?

8.50 Un fabricante de cierta marca de barras de cereal con bajo contenido de grasa afirma que el contenido promedio de grasa saturada en éstas es de 0.5 gramos. En una muestra aleatoria de 8 barras de cereal de esta marca se encontró que su contenido de grasa saturada era de 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4 y 0.2. ¿Estaría de acuerdo con tal afirmación? Suponga una distribución normal.

8.51 Para una distribución F calcule:
a) $f_{0.05}$ con $v_1 = 7$ y $v_2 = 15$;
b) $f_{0.05}$ con $v_1 = 15$ y $v_2 = 7$;
c) $f_{0.01}$ con $v_1 = 24$ y $v_2 = 19$;
d) $f_{0.95}$ con $v_1 = 19$ y $v_2 = 24$;
e) $f_{0.99}$ con $v_1 = 28$ y $v_2 = 12$.

8.52 Se aplican pruebas a 10 cables conductores soldados a un dispositivo semiconductor con el fin de determinar su resistencia a la tracción. Las pruebas demostraron que para romper la unión se requieren las libras de fuerza que se listan a continuación:

19.8	12.7	13.2	16.9	10.6
18.8	11.1	14.3	17.0	12.5

Otro conjunto de 8 cables conductores que forman un dispositivo se encapsuló y se probó para determinar si el encapsulado aumentaba la resistencia a la tracción. Las pruebas dieron los siguientes resultados:

24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5

Comente acerca de la evidencia disponible respecto a la igualdad de las dos varianzas de población.

8.53 Considere las siguientes mediciones de la capacidad de producción de calor del carbón producido por dos minas (en millones de calorías por tonelada):

Mina 1:	8260	8130	8350	8070	8340	
Mina 2:	7950	7890	7900	8140	7920	7840

¿Se puede concluir que las dos varianzas de población son iguales?

8.54 Dibuje una gráfica de cuantiles con los siguientes datos, que representan la vida, en horas, de cincuenta lámparas incandescentes esmeriladas de 40 watts y 110 voltios, tomados de pruebas de vida forzadas:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

8.55 Dibuje una gráfica de cuantiles-cuantiles normales con los siguientes datos, que representan los diámetros de 36 cabezas de remache en 1/100 de una pulgada:

6.72	6.77	6.82	6.70	6.78	6.70	6.62
6.75	6.66	6.66	6.64	6.76	6.73	6.80
6.72	6.76	6.76	6.68	6.66	6.62	6.72
6.76	6.70	6.78	6.76	6.67	6.70	6.72
6.74	6.81	6.79	6.78	6.66	6.76	6.76
6.72						

Ejercicios de repaso

8.56 Considere los datos que se presentan en el ejercicio 1.20 de la página 31. Dibuje una gráfica de caja y extensión, y comente acerca de la naturaleza de la muestra. Calcule la media muestral y la desviación estándar de la muestra.

8.57 Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes que tienen distribuciones exponenciales idénticas con parámetro θ, demuestre que la función de densidad de la variable aleatoria $Y = X_1 + X_2 + \cdots + X_n$ es la de una distribución gamma con parámetros $\alpha = n$ y $\beta = \theta$.

8.58 Al probar el monóxido de carbono que contiene cierta marca de cigarrillos, los datos que se obtuvieron, en miligramos por cigarrillo, se codificaron restando 12 a cada observación. Utilice los resultados del ejercicio 8.14 de la página 231 para calcular la desviación estándar del contenido de monóxido de carbono de una muestra aleatoria de 15 cigarrillos de esta marca, si las mediciones codificadas son 3.8, −0.9, 5.4, 4.5, 5.2, 5.6, −0.1, −0.3, −1.7, 5.7, 3.3, 4.4, −0.5 y 1.9.

8.59 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaños $n_1 = 8$ y $n_2 = 12$, tomadas de poblaciones normales con varianzas iguales, calcule $P(S_1^2 / S_2^2 < 4.89)$.

8.60 Una muestra aleatoria de 5 presidentes de bancos indicó sueldos anuales de $395,000, $521,000, $483,000, $479,000 y $510,000. Calcule la varianza de este conjunto.

8.61 Si el número de huracanes que azotan cierta área del este de Estados Unidos cada año es una variable aleatoria que tiene una distribución de Poisson con $\mu = 6$, calcule la probabilidad de que esta área sea azotada por
a) exactamente 15 huracanes en 2 años;
b) a lo sumo 9 huracanes en 2 años.

8.62 Una empresa de taxis prueba una muestra aleatoria de 10 neumáticos radiales con bandas tensoras de acero de cierta marca y registra los siguientes desgastes de la banda: 48,000, 53,000, 45,000, 61,000, 59,000, 56,000, 63,000, 49,000, 53,000 y 54,000 kilómetros.

Utilice los resultados del ejercicio 8.14 de la página 231 para calcular la desviación estándar de este conjunto de datos dividiendo primero cada observación entre 1000 y después restando 55 al resultado.

8.63 Considere los datos del ejercicio 1.19 de la página 31. Dibuje una gráfica de caja y extensión. Comente y calcule la media muestral y la desviación estándar muestral.

8.64 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaños $n_1 = 25$ y $n_2 = 31$, tomadas de poblaciones normales con varianzas $\sigma_1^2 = 10$ y $\sigma_2^2 = 15$, respectivamente, calcule

$$P(S_1^2/S_2^2 > 1.26).$$

8.65 Considere el ejemplo 1.5 de la página 25. Comente acerca de cualquier valor extremo.

8.66 Considere el ejercicio de repaso 8.56. Comente acerca de cualquier valor extremo en los datos.

8.67 La resistencia a la rotura X de cierto remache que se utiliza en el motor de una máquina tiene una media de 5000 psi y una desviación estándar de 400 psi. Se toma una muestra aleatoria de 36 remaches. Considere la distribución de $\bar{X}$, la media muestral de la resistencia a la rotura.

a) ¿Cuál es la probabilidad de que la media de la muestra caiga entre 4800 psi y 5200 psi?

b) ¿Qué muestra n sería necesaria para tener

$$P(4900 < \bar{X} < 5100) = 0.99?$$

8.68 Considere la situación del ejercicio de repaso 8.62. Si la población de la cual se tomó la muestra tiene una media poblacional $\mu = 53{,}000$ kilómetros, ¿esta información de la muestra parece apoyar esa afirmación? En su respuesta calcule

$$t = \frac{\bar{x} - 53{,}000}{s/\sqrt{10}}$$

y determine, consultando la tabla A.4 (con 9 g.l.), si el valor t calculado es razonable o si parece ser un suceso raro.

8.69 Se consideran dos propulsores de combustible sólido distintos, el tipo A y el tipo B, para una actividad del programa espacial. Las velocidades de combustión en el propulsor son fundamentales. Se toman muestras aleatorias de 20 especímenes de los dos propulsores con medias muestrales de 20.5 cm/s para el propulsor A y de 24.50 cm/s para el propulsor B. Por lo general se supone que la variabilidad en la velocidad de combustión es casi igual para los dos propulsores y que es determinada por una desviación estándar de población de 5 cm/s. Suponga que la velocidad de combustión para cada propulsor es aproximadamente normal, por lo cual se debería utilizar el teorema del límite central. Nada se sabe acerca de las medias poblacionales de las dos velocidades de combustión y se espera que este experimento revele algo sobre ellas.

a) Si, de hecho, $\mu_A = \mu_B$, ¿cuál será $P(\bar{X}_B - \bar{X}_A \geq 4.0)$?

b) Utilice lo que respondió en el inciso *a*) para dar luz sobre la validez de la proposición $\mu_A = \mu_B$.

8.70 La concentración de un ingrediente activo en el producto de una reacción química es fuertemente influido por el catalizador que se usa en la reacción. Se considera que cuando se utiliza el catalizador A la concentración media de la población excede el 65%. Se sabe que la desviación estándar es $\sigma = 5\%$. Una muestra de productos tomada de 30 experimentos independientes proporciona la concentración promedio de $\bar{x}_A = 64.5\%$.

a) ¿Esta información muestral, con una concentración promedio de $\bar{x}_A = 64.5\%$, ofrece información inquietante de que quizá μ_A no sea el 65% sino menos que ese porcentaje? Respalde su respuesta con una aseveración de probabilidad.

b) Suponga que se realiza un experimento similar utilizando otro catalizador, el B. Se supone que la desviación estándar σ sigue siendo 5% y $\bar{x}_B$ resulta ser 70%. Comente si la información muestral del catalizador B sugiere con certeza que μ_B es en realidad mayor que μ_A. Respalde su respuesta calculando

$$P(\bar{X}_B - \bar{X}_A \geq 5.5 \mid \mu_B = \mu_A).$$

c) En el caso de que $\mu_A = \mu_B = 65\%$, determine la distribución aproximada de las siguientes cantidades (con la media y la varianza de cada una). Utilice el teorema del límite central.

i) $\bar{X}_B$;
ii) $\bar{X}_A - \bar{X}_B$;
iii) $\frac{\bar{X}_A - \bar{X}_B}{\sigma\sqrt{2/30}}$.

8.71 Con la información del ejercicio de repaso 8.70 calcule (suponiendo $\mu_B = 65\%$) $P(\bar{X}_B \geq 70)$.

8.72 Dada una variable aleatoria normal X con media 20 y varianza 9, y una muestra aleatoria de tamaño n tomada de la distribución, ¿qué tamaño de la muestra n se necesita para que

$$P(19.9 \leq \bar{X} \leq 20.1) = 0.95?$$

8.73 En el capítulo 9 se estudiará con detenimiento el concepto de **estimación de parámetros**. Suponga que X es una variable aleatoria con media μ y varianza $\sigma^2 = 1.0$. Además, suponga que se toma una muestra aleato-

ria de tamaño n y que $\bar{x}$ se utiliza como un *estimado* de μ. Cuando se toman los datos y se mide la media de la muestra, deseamos que ésta esté dentro de 0.05 unidades de la media real con una probabilidad de 0.99. Es decir, aquí queremos que haya muchas posibilidades de que la $\bar{x}$ calculada de la muestra esté "muy cerca de" la media de población (¡dondequiera que ésta se encuentre!), de manera que deseamos

$$P(|\bar{X} - \mu| > 0.05) = 0.99.$$

¿Qué tamaño de muestra se requiere?

8.74 Suponga que se utiliza una máquina para llenar envases de cartón con un líquido. La especificación que es estrictamente indispensable para el llenado de la máquina es 9 ± 1.5 onzas. El proveedor considera que cualquier envase de cartón que no cumpla con tales límites de peso en el llenado está defectuoso. Se espera que al menos 99% de los envases de cartón cumplan con la especificación. En el caso de que $\mu = 9$ y $\sigma = 1$, ¿qué proporción de envases de cartón del proceso están defectuosos? Si se hacen cambios para reducir la variabilidad, ¿cuánto se tiene que reducir σ para que haya 0.99 de probabilidades de cumplir con la especificación? Suponga una distribución normal para el peso.

8.75 Considere la situación del ejercicio de repaso 8.74. Suponga que se hace un gran esfuerzo para "estrechar" la variabilidad del sistema. Después de eso se toma una muestra aleatoria de tamaño 40 de la nueva línea de ensamble y se obtiene que la varianza de la muestra es $s^2 = 0.188$ onzas^2. ¿Tenemos evidencia numérica sólida de que σ^2 se redujo a menos de 1.0? Considere la probabilidad

$$P(S^2 \leq 0.188 \mid \sigma^2 = 1.0),$$

y dé una conclusión.

8.76 **Proyecto de grupo:** Divida al grupo en equipos de cuatro estudiantes. Cada equipo deberá ir al gimnasio de la universidad o a un gimnasio local y preguntar a cada persona que cruce el umbral cuánto mide en pulgadas. Después, cada equipo dividirá los datos de las estaturas por género y trabajará en conjunto para realizar las actividades que se indican a continuación.

a) Dibujen una gráfica de cuantiles-cuantiles normal con los datos. Si usan la gráfica como base, ¿les parecería que los datos tienen una distribución normal?

b) Utilicen la varianza muestral como un estimado de la varianza real para cada género. Supongan que la estatura media de la población de los hombres es realmente tres pulgadas más grande que la de las mujeres. ¿Cuál es la probabilidad de que la estatura promedio de los hombres sea 4 pulgadas más grande que la de las mujeres en su muestra?

c) ¿Qué factores podrían provocar que estos resultados sean engañosos?

8.9 Posibles riesgos y errores conceptuales. Relación con el material de otros capítulos

El teorema del límite central es una de las más poderosas herramientas de la estadística, y aunque este capítulo es relativamente breve, contiene gran cantidad de información fundamental acerca de las herramientas que se utilizarán en el resto del libro.

El concepto de distribución muestral es una de las ideas fundamentales más importantes de la estadística y, en este momento de su entrenamiento, el estudiante debería entenderlo con claridad antes de continuar con los siguientes capítulos, en los cuales se continuarán utilizando ampliamente las distribuciones muestrales. Suponga que se quiere utilizar el estadístico $\bar{X}$ para hacer inferencias acerca de la media de la población μ, lo cual se hace utilizando el valor observado $\bar{x}$ de una sola muestra de tamaño n. Luego, cualquier inferencia deberá hacerse tomando en cuenta no sólo el valor único, sino también la estructura teórica o la **distribución de todos los valores $\bar{x}$ que se podrían observar a partir de las muestras de tamaño n**. Como resultado de lo anterior surge el concepto de *distribución muestral*, que es la base del teorema del límite central. Las distribuciones t, χ^2 y F también se utilizan en el contexto de las distribuciones muestrales. Por ejemplo, la distribución t, que se ilustra en la figura 8.8, representa la estructura que ocurre si se forman todos los valores de $\frac{\bar{x} - \mu}{s/\sqrt{n}}$, donde $\bar{x}$ y s se toman de las

muestras de tamaño n de una distribución $n(x; \mu, \sigma)$. Se pueden hacer comentarios similares en relación con χ^2 y F, y el lector no debería olvidar que la información muestral que conforma los estadísticos para todas estas distribuciones es la normal. Por lo tanto, se podría afirmar que **donde haya una *t*, *F* o χ^2 la fuente era una muestra de una distribución normal**.

Podría parecer que las tres distribuciones antes descritas se presentaron de una forma bastante aislada, sin indicar a qué se refieren. Sin embargo, aparecerán en la resolución de problemas prácticos a lo largo del texto.

Ahora bien, hay tres cuestiones que se deben tener presentes para evitar que haya confusión respecto a estas distribuciones muestrales fundamentales:

i) No se puede usar el teorema del límite central a menos que se conozca σ. Para usar el teorema del límite central cuando no se conoce σ se debe reemplazar con s, la desviación estándar de la muestra.

ii) El estadístico T **no** es un resultado del teorema del límite central y $x_1, x_2, ..., x_n$ deben provenir de una distribución $n(x; \mu, \sigma)$ para que $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ sea una distribución t; por supuesto, s es tan sólo una estimación de σ.

iii) Aunque el concepto de **grados de libertad** es nuevo en este punto, debería ser muy intuitivo, ya que es razonable que la naturaleza de la distribución de S y también t deban depender de la cantidad de información en la muestra $x_1, x_2, ..., x_n$.

Capítulo 9

Problemas de estimación de una y dos muestras

9.1 Introducción

En los capítulos anteriores destacamos las propiedades del muestreo de la media y de la varianza muestrales. También destacamos las representaciones de datos en varias formas. El propósito de estas presentaciones es establecer las bases que permitan a los estadísticos sacar conclusiones acerca de los parámetros de poblaciones tomadas de datos experimentales. Por ejemplo, el teorema del límite central brinda información sobre la distribución de la media muestral $\bar{X}$. La distribución incluye la media de la población μ. Por consiguiente, cualesquiera conclusiones respecto a μ, extraídas de un promedio muestral observado, deben depender de lo que se sabe acerca de su distribución muestral. Se podría decir algo similar en lo que se refiere a S^2 y σ^2. Como es evidente, es muy probable que cualquier conclusión que saquemos acerca de la varianza de una distribución normal implique la distribución muestral de S^2.

En este capítulo comenzaremos por presentar de manera formal el propósito de la inferencia estadística. Continuaremos con el análisis del problema de la **estimación de los parámetros de la población**. Restringiremos nuestros desarrollos formales de los procedimientos de estimación específicos a problemas que impliquen una y dos muestras.

9.2 Inferencia estadística

En el capítulo 1 presentamos la filosofía general de la inferencia estadística formal. La **inferencia estadística** consta de los métodos mediante los cuales se hacen inferencias o generalizaciones acerca de una población. La tendencia actual es distinguir entre el **método clásico** de estimación de un parámetro de la población, donde las inferencias se basan estrictamente en información obtenida de una muestra aleatoria seleccionada de la población, y el **método bayesiano**, el cual utiliza el conocimiento subjetivo que ya se posee sobre la distribución de probabilidad de los parámetros desconocidos junto con la información que proporcionan los datos de la muestra. En la mayor parte de este capítulo utilizaremos los métodos clásicos para estimar los parámetros de la población desconocidos, como la media, la proporción y la varianza, mediante el cálculo de estadísticos de muestras aleatorias y la aplicación de la teoría de las distribuciones muestrales, gran

parte de lo cual se estudió en el capítulo 8. La estimación bayesiana se analizará en el capítulo 18.

La inferencia estadística se puede dividir en dos áreas principales: **estimación** y **pruebas de hipótesis**. Trataremos estas dos áreas por separado: en este capítulo veremos la teoría y las aplicaciones de la estimación, y en el capítulo 10 revisaremos la prueba de hipótesis. Para distinguir claramente un área de la otra, considere los siguientes ejemplos. Un candidato a un cargo público podría estar interesado en estimar la verdadera proporción de votantes que lo favorecerán mediante la obtención de las opiniones de una muestra aleatoria de 100 de ellos. La parte de votantes en la muestra que favorecerán al candidato se podría utilizar como un estimado de la verdadera proporción en la población de votantes. El conocimiento de la distribución muestral de una proporción nos permite establecer el grado de exactitud de tal estimado. Este problema cae en el área de la estimación.

Considere ahora el caso de alguien a quien le interesa averiguar si la marca *A* de cera para piso es más resistente al desgaste que la marca *B*. Se podría plantear la hipótesis de que la marca *A* es mejor que la marca *B* y, después de la prueba adecuada, aceptar o rechazar dicha hipótesis. En este ejemplo no intentamos estimar un parámetro, sino llegar a una decisión correcta acerca de una hipótesis planteada previamente. Una vez más, dependemos de la teoría del muestreo y de utilizar datos que nos proporcionen alguna medida del grado de exactitud de nuestra decisión.

9.3 Métodos de estimación clásicos

La **estimación puntual** de algún parámetro de la población θ es un solo valor $\hat{\theta}$ de un estadístico $\hat{\Theta}$. Por ejemplo, el valor $\bar{x}$ del estadístico $\bar{X}$, que se calcula a partir de una muestra de tamaño n, es una estimación puntual del parámetro de la población μ. De manera similar, $\hat{p} = x/n$ es una estimación puntual de la verdadera proporción p para un experimento binomial.

No se espera que un estimador logre estimar el parámetro de la población sin error. No se espera que $\bar{X}$ estime μ con exactitud, lo que en realidad se espera es que no esté muy alejada. Para una muestra específica, la manera en que se podría obtener un estimado más cercano de μ es utilizando la mediana de la muestra $\tilde{X}$ como estimador. Considere, por ejemplo, una muestra que consta de los valores 2, 5 y 11 de una población cuya media es 4, la cual, supuestamente, se desconoce. Podríamos estimar μ para que sea $\bar{x} = 6$ usando la media muestral como nuestro estimado, o bien, $\tilde{x} = 5$ utilizando la mediana muestral. En este caso el estimador $\tilde{X}$ produce una estimación más cercana al parámetro verdadero que la que produce el estimador $\bar{X}$. Por otro lado, si nuestra muestra aleatoria contiene los valores 2, 6 y 7, entonces $\bar{x} = 5$ y $\tilde{x} = 6$, de manera que el mejor estimador es $\bar{X}$. Cuando no conocemos el valor real de μ, tenemos que comenzar por decidir qué estimador utilizaremos, si $\bar{X}$ o $\tilde{X}$.

Estimador insesgado

¿Cuáles son las propiedades que una "buena" función de decisión debería tener para poder influir en nuestra elección de un estimador en vez de otro? Sea $\hat{\Theta}$ un estimador cuyo valor $\hat{\theta}$ es una estimación puntual de algún parámetro de la población desconocido θ. Sin duda desearíamos que la distribución muestral de $\hat{\Theta}$ tuviera una media igual al parámetro estimado. Al estimador que tuviera esta propiedad se le llamaría **estimador insesgado**.

Definición 9.1: Se dice que un estadístico $\hat{\Theta}$ es un **estimador insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Ejemplo 9.1: Demuestre que S^2 es un estimador insesgado del parámetro σ^2.

***Solución*:** En la sección 8.5, en la página 244, demostramos que

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Entonces,

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma_{X_i}^2 - n\sigma_{\bar{X}}^2\right).$$

Sin embargo,

$$\sigma_{X_i}^2 = \sigma^2, \text{ para } i = 1, 2, \ldots, n, \text{ y } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Por lo tanto,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Aunque S^2 es un estimador insesgado de σ^2, S, por otro lado, suele ser un estimador sesgado de σ, un sesgo que en el caso de muestras grandes se vuelve insignificante. Este ejemplo ilustra **por qué dividimos entre $n - 1$** en vez de entre n cuando estimamos la varianza.

Varianza de un estimador puntual

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro de la población θ, deseamos elegir el estimador cuya distribución muestral tenga la menor varianza. Por lo tanto, si $\sigma_{\hat{\theta}_1}^2 < \sigma_{\hat{\theta}_2}^2$, decimos que $\hat{\Theta}_1$ es un **estimador más eficaz** de θ que $\hat{\Theta}_2$.

Definición 9.2: Si consideramos todos los posibles estimadores insesgados de algún parámetro θ, al que tiene la menor varianza lo llamamos estimador más eficaz de θ.

En la figura 9.1 se ilustran las distribuciones muestrales de tres estimadores diferentes $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$, todos para θ. Es evidente que sólo $\hat{\Theta}_1$ y $\hat{\Theta}_2$ no son sesgados, ya que sus distribuciones están centradas en θ. El estimador $\hat{\Theta}_1$ tiene una varianza menor que $\hat{\Theta}_2$, por lo tanto, es más eficaz. En consecuencia, el estimador de θ que elegiríamos, entre los tres que estamos considerando, sería $\hat{\Theta}_1$.

Para poblaciones normales se puede demostrar que tanto $\bar{X}$ como $\tilde{X}$ son estimadores insesgados de la media de la población μ, pero la varianza de $\bar{X}$ es más pequeña que la varianza de $\tilde{X}$. Por consiguiente, los estimados $\bar{x}$ y $\tilde{x}$ serán, en promedio, iguales a

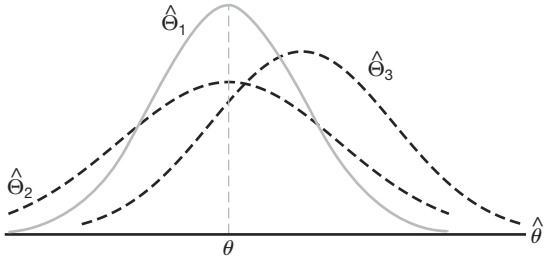


Figura 9.1: Distribuciones muestrales de diferentes estimadores de θ.

la media de la población μ, aunque podría ser que $\bar{x}$ esté más cerca de μ para una muestra dada y, por lo tanto, que $\tilde{X}$ sea más eficaz que $\bar{X}$.

Estimación por intervalo

Podría ser que ni el estimador insesgado más eficaz estime con exactitud el parámetro de la población. Es cierto que la exactitud de la estimación aumenta cuando las muestras son grandes; pero incluso así no tenemos razones para esperar que una **estimación puntual** de una muestra dada sea exactamente igual al parámetro de la población que se supone debe estimar. Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se conoce como **estimación por intervalo**.

Una estimación por intervalo de un parámetro de la población θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$. Por ejemplo, una muestra aleatoria de calificaciones verbales de la prueba SAT para estudiantes universitarios de primer año produciría un intervalo de 530 a 550, dentro del cual esperamos encontrar el promedio verdadero de todas las calificaciones verbales de la prueba SAT para ese grupo. Los valores de los puntos extremos, 530 y 550, dependerán de la media muestral calculada $\bar{x}$ y de la distribución de muestreo de $\bar{X}$. A medida que aumenta el tamaño de la muestra, sabemos que $\sigma_{\bar{X}}^2 = \sigma^2/n$ disminuye y, en consecuencia, cabe la posibilidad de que nuestra estimación se acerque más al parámetro μ, lo cual daría como resultado un intervalo más corto. De esta manera, el intervalo de la estimación indica, por su longitud, la precisión de la estimación puntual. Un ingeniero obtendrá información acerca de la proporción de la población de artículos defectuosos tomando una muestra y calculando la *proporción muestral defectuosa,* sin embargo, una estimación por intervalo podría ser más informativa.

Interpretación de las estimaciones por intervalo

Como muestras distintas suelen producir valores diferentes de $\hat{\Theta}$ y, por lo tanto, valores diferentes de $\hat{\theta}_L$ y $\hat{\theta}_U$, estos puntos extremos del intervalo son valores de las variables aleatorias correspondientes $\hat{\Theta}_L$ y $\hat{\Theta}_U$. De la distribución muestral de $\hat{\Theta}$ seremos capaces de determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ sea igual a cualquier

valor positivo de una fracción que queramos especificar. Si, por ejemplo, calculamos $\hat{\Theta}_L$ y $\hat{\Theta}_U$, tales que

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

para $0 < \alpha < 1$, tenemos entonces una probabilidad de $1 - \alpha$ de seleccionar una muestra aleatoria que produzca un intervalo que contenga θ. El intervalo $\hat{\theta}_L < \theta < \hat{\theta}_U$, que se calcula a partir de la muestra seleccionada, se llama entonces **intervalo de confianza** del $100(1 - \alpha)\%$, la fracción $1 - \alpha$ se denomina **coeficiente de confianza** o **grado de confianza**, y los extremos, $\hat{\theta}_L$ y $\hat{\theta}_U$, se denominan **límites de confianza** inferior y superior. Así, cuando $\alpha = 0.05$, tenemos un intervalo de confianza del 95%, y cuando $\alpha = 0.01$ obtenemos un intervalo de confianza más amplio del 99%. Cuanto más amplio sea el intervalo de confianza, más confiaremos en que contiene el parámetro desconocido. Desde luego, es mejor tener un 95% de confianza en que la vida promedio de cierto transistor de un televisor está entre los 6 y los 7 años, que tener un 99% de confianza en que esté entre los 3 y los 10 años. De manera ideal, preferimos un intervalo corto con un grado de confianza alto. Algunas veces las restricciones en el tamaño de nuestra muestra nos impiden tener intervalos cortos sin sacrificar cierto grado de confianza.

En las siguientes secciones estudiaremos los conceptos de estimación puntual y por intervalos, y en cada sección presentaremos un caso especial diferente. El lector debería notar que, aunque la estimación puntual y por intervalos representan diferentes aproximaciones para obtener información respecto a un parámetro, están relacionadas debido a que los estimadores del intervalo de confianza se basan en estimadores puntuales. En la siguiente sección, por ejemplo, veremos que $\bar{X}$ es un estimador puntual de μ muy razonable. Como resultado, el importante estimador del intervalo de confianza de μ depende del conocimiento de la distribución muestral de $\bar{X}$.

Empezaremos la siguiente sección con el caso más sencillo de un intervalo de confianza, en donde el escenario es simple pero poco realista. Nos interesa estimar una media de la población μ cuando σ todavía se desconoce. Evidentemente, si se desconoce μ es muy improbable que se conozca σ. Cualquier información histórica que produzca datos suficientes para permitir suponer que se conoce σ probablemente habría producido información similar acerca de μ. A pesar de este argumento iniciamos con este caso porque los conceptos y los mecanismos resultantes asociados con la estimación del intervalo de confianza también estarán asociados con las situaciones más realistas que presentaremos más adelante en la sección 9.4 y las siguientes.

9.4 Una sola muestra: estimación de la media

La distribución muestral de $\bar{X}$ está centrada en μ y en la mayoría de las aplicaciones la varianza es más pequeña que la de cualesquiera otros estimadores de μ. Por lo tanto, se utilizará la media muestral $\bar{x}$ como una estimación puntual para la media de la población μ. Recuerde que $\sigma^2_{\bar{X}} = \sigma^2/n$, por lo que una muestra grande producirá un valor de $\bar{X}$ procedente de una distribución muestral con varianza pequeña. Por consiguiente, es probable que $\bar{x}$ sea una estimación muy precisa de μ cuando n es grande.

Consideremos ahora la estimación por intervalos de μ. Si seleccionamos nuestra muestra a partir de una población normal o, a falta de ésta, si n es suficientemente grande, podemos establecer un intervalo de confianza para μ considerando la distribución muestral de $\bar{X}$.

De acuerdo con el teorema del límite central, podemos esperar que la distribución muestral de $\bar{X}$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Al escribir $z_{\alpha/2}$ para el valor z por arriba del cual encontramos una área de $\alpha/2$ bajo la curva normal, en la figura 9.2 podemos ver que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

En consecuencia,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

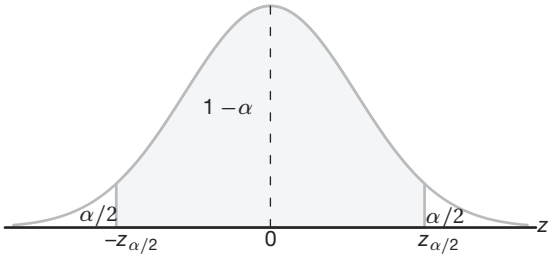


Figura 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Si multiplicamos cada término en la desigualdad por $\sigma/\sqrt{n}$ y después restamos $\bar{X}$ de cada término, y en seguida multiplicamos por − 1 (para invertir el sentido de las desigualdades), obtenemos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se selecciona una muestra aleatoria de tamaño n de una población cuya varianza σ^2 se conoce y se calcula la media $\bar{x}$ para obtener el intervalo de confianza $100(1 - \alpha)\%$. Es importante enfatizar que recurrimos al teorema del límite central citado anteriormente. Como resultado, es importante observar las condiciones para las aplicaciones que siguen.

Intervalo de confianza de μ cuando se conoce σ^2

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población de la que se conoce su varianza σ^2, lo que da un intervalo de confianza de $100(1 - \alpha)\%$ para μ es

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

En el caso de muestras pequeñas que se seleccionan de poblaciones no normales, no podemos esperar que nuestro grado de confianza sea preciso. Sin embargo, para muestras

de tamaño $n \geq 30$, en las que la forma de las distribuciones no esté muy sesgada, la teoría de muestreo garantiza buenos resultados.

Queda claro que los valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$, las cuales se definieron en la sección 9.3, son los límites de confianza

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{y} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Muestras diferentes producirán valores diferentes de $\bar{x}$ y, por lo tanto, producirán diferentes estimaciones por intervalos del parámetro μ, como se muestra en la figura 9.3. Los puntos en el centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Observe que todos los intervalos tienen el mismo ancho, pues esto depende sólo de la elección de $z_{\alpha/2}$ una vez que se determina $\bar{x}$. Cuanto más grande sea el valor de $z_{\alpha/2}$ que elijamos, más anchos haremos todos los intervalos, y podremos tener más confianza en que la muestra particular que seleccionemos producirá un intervalo que contenga el parámetro desconocido μ. En general, para una elección de $z_{\alpha/2}$, $100(1-\alpha)\%$ de los intervalos contendrá μ.

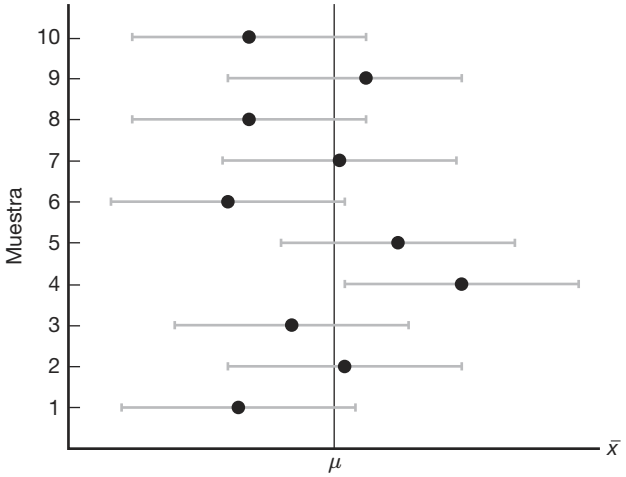


Figura 9.3: Estimaciones por intervalos de μ para muestras diferentes.

Ejemplo 9.2: Se encuentra que la concentración promedio de zinc que se obtiene en una muestra de mediciones en 36 sitios diferentes de un río es de 2.6 gramos por mililitro. Calcule los intervalos de confianza del 95% y 99% para la concentración media de zinc en el río. Suponga que la desviación estándar de la población es de 0.3 gramos por mililitro.

***Solución*:** La estimación puntual de μ es $\bar{x} = 2.6$. El valor z que deja una área de 0.025 a la derecha y, por lo tanto, una área de 0.975 a la izquierda es $z_{0.025} = 1.96$ (véase la tabla A.3). En consecuencia, el intervalo de confianza del 95% es

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

que se reduce a $2.50 < \mu < 2.70$. Para calcular un intervalo de confianza del 99% encontramos el valor z que deja una área de 0.005 a la derecha y de 0.995 a la izquierda. Por lo tanto, usando la tabla A.3 nuevamente, $z_{0.005} = 2.575$ y el intervalo de confianza de 99% es

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right),$$

o simplemente

$$2.47 < \mu < 2.73.$$

Ahora vemos que se requiere un intervalo más grande para estimar μ con un mayor grado de confianza. ∎

El intervalo de confianza del $100(1-\alpha)\%$ ofrece un estimado de la precisión de nuestra estimación puntual. Si μ es realmente el valor central del intervalo, entonces $\bar{x}$ estima μ sin error. La mayoría de las veces, sin embargo, $\bar{x}$ no será exactamente igual a μ y la estimación puntual será errónea. La magnitud de este error será el valor absoluto de la diferencia entre μ y $\bar{x}$, de manera que podemos tener $100(1-\alpha)\%$ de confianza en que esta diferencia no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Podemos ver esto fácilmente dibujando un diagrama de un intervalo de confianza hipotético, como el de la figura 9.4.

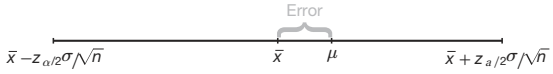


Figura 9.4: Error en la estimación de μ mediante $\bar{x}$.

Teorema 9.1: Si utilizamos $\bar{x}$ como una estimación de μ, podemos tener $100(1-\alpha)\%$ de confianza en que el error no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

En el ejemplo 9.2 tenemos una confianza del 95% en que la media muestral $\bar{x} = 2.6$ difiere de la media verdadera μ en una cantidad menor que $(1.96)(0.3)/\sqrt{36} = 0.1$ y 99% de confianza en que la diferencia es menor que $(2.575)(0.3)/\sqrt{36} = 0.13$.

Con frecuencia queremos saber qué tan grande necesita ser una muestra para poder estar seguros de que el error al estimar μ será menor que una cantidad específica e. Por medio del teorema 9.1 debemos elegir n de manera que $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e$. Al resolver esta ecuación obtenemos la siguiente fórmula para n.

Teorema 9.2: Si usamos $\bar{x}$ como una estimación de μ, podemos tener $100(1-\alpha)\%$ de confianza en que el error no excederá a una cantidad específica e cuando el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Cuando resolvemos para la muestra con tamaño n, redondeamos todos los valores decimales al siguiente número entero. Si seguimos este principio, podemos estar seguros de que nuestro grado de confianza nunca caerá por debajo del $100(1-\alpha)\%$.

En términos estrictos, la fórmula del teorema 9.2 sólo será aplicable si se conoce la varianza de la población de la cual se seleccionó la muestra. Si no contamos con esa información, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ para proporcionar una estimación de σ. Después, usando s como aproximación para σ en el teorema 9.2, podemos determinar aproximadamente cuántas observaciones necesitamos para brindar el grado de precisión deseado.

Ejemplo 9.3: ¿Qué tan grande debe ser la muestra del ejemplo 9.2 si queremos tener 95% de confianza en que nuestra estimación de μ diferirá por menos de 0.05?

***Solución*:** La desviación estándar de la población es $\sigma = 0.3$. Entonces, por medio del teorema 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

Por lo tanto, podemos tener 95% de confianza en que una muestra aleatoria de tamaño 139 proporcionará una estimación $\bar{x}$ que diferirá de μ en una cantidad menor que 0.05. ∎

Límites de confianza unilaterales

Los intervalos de confianza y los límites de confianza resultantes que hasta ahora hemos analizado en realidad son *bilaterales,* es decir, tienen límites superior e inferior. Sin embargo, hay muchas aplicaciones en las que sólo se requiere un límite. Por ejemplo, si a un ingeniero le interesara determinar una medida de resistencia a la tensión, la información que más le ayudaría a lograr su objetivo sería la del límite inferior, ya que éste indica el escenario del "peor caso", es decir, el de la menor resistencia. Por otro lado, si se buscara determinar una medida para la cual un valor de μ relativamente grande no fuera redituable o deseable, entonces la medida que resultaría de interés sería la del límite de confianza superior. Un ejemplo en el que la medida del límite superior sería muy informativa es el caso en el que se necesita hacer inferencias para determinar la composición media de mercurio en el agua de un río.

Los límites de confianza unilaterales se desarrollan de la misma forma que los intervalos bilaterales. Sin embargo, la fuente es un enunciado de probabilidad unilateral que utiliza el teorema del límite central:

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

Entonces, es posible manipular el enunciado de probabilidad de forma muy similar a como se hizo anteriormente para obtener

$$P(\mu > \bar{X} - z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Una manipulación similar de $P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$ da

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Como resultado, se obtienen los siguientes límites unilaterales superior e inferior.

Límites de confianza unilaterales de μ cuando se conoce el valor de σ^2

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n a partir de una población con varianza σ^2, los límites de confianza unilaterales del $100(1 - \alpha)\%$ para μ son dados por

límite unilateral superior: $\bar{x} + z_\alpha \sigma/\sqrt{n}$;

límite unilateral inferior: $\bar{x} - z_\alpha \sigma/\sqrt{n}$.

Ejemplo 9.4: En un experimento de pruebas psicológicas se seleccionan al azar 25 sujetos y se miden sus tiempos de reacción, en segundos, ante un estímulo particular. La experiencia sugiere que la varianza en los tiempos de reacción ante los diferentes tipos de estímulos es de 4 s^2 y que la distribución del tiempo de reacción es aproximadamente normal. El tiempo promedio para los sujetos fue de 6.2 segundos. Calcule un límite superior del 95% para el tiempo medio de reacción.

***Solución*:** Lo que da el límite superior del 95% es

$$\begin{aligned}\bar{x} + z_\alpha \sigma/\sqrt{n} &= 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658 \\ &= 6.858 \text{ segundos.}\end{aligned}$$

En consecuencia, tenemos un 95% de confianza en que el tiempo promedio de reacción es menor que 6.858 segundos. ∎

El caso en que se desconoce σ

Con frecuencia debemos tratar de estimar la media de una población sin conocer la varianza. El lector debería recordar que en el capítulo 8 aprendió que, si tenemos una muestra aleatoria a partir de una *distribución normal,* entonces la variable aleatoria

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución *t* de Student con $n - 1$ grados de libertad. Aquí *S* es la desviación estándar de la muestra. En esta situación, en la que se desconoce σ, se puede utilizar *T* para construir un intervalo de confianza para μ. El procedimiento es igual que cuando se conoce σ, sólo que en este caso σ se reemplaza con *S* y la distribución normal estándar se reemplaza con la distribución *t*. Si nos remitimos a la figura 9.5, podemos afirmar que

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor *t* con $n - 1$ grados de libertad, por arriba del cual encontramos una área de $\alpha/2$. Debido a la simetría, un área igual de $\alpha/2$ caerá a la izquierda de $-t_{\alpha/2}$. Al sustituir por *T* escribimos

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Al multiplicar cada término en la desigualdad por $S/\sqrt{n}$ y después restar $\bar{X}$ de cada término y multiplicar por –1, obtenemos

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño *n* se calculan la media $\bar{x}$ y la desviación estándar *s,* y se obtiene el siguiente intervalo de confianza $100(1 - \alpha)\%$ para μ.

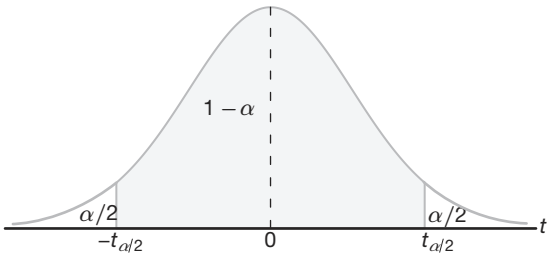


Figura 9.5: $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$.

Intervalo de confianza para μ cuando se desconoce σ^2

Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población normal de la que se desconoce la varianza σ^2, un intervalo de confianza del $100(1-\alpha)\%$ para μ es

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha.

Hicimos una distinción entre los casos en los que se conoce σ y en los que se desconoce calculando las estimaciones del intervalo de confianza. Deberíamos resaltar que para el caso en que se conoce σ se utiliza el teorema del límite central, mientras que, para el caso en que se desconoce, se usa la distribución muestral de la variable aleatoria T. Sin embargo, el uso de la distribución t se basa en la premisa de que el muestreo es de una distribución normal. Siempre que la forma de la distribución se aproxime a la de campana, se puede utilizar la distribución t para calcular los intervalos de confianza cuando se desconoce σ^2, y se pueden esperar muy buenos resultados.

Los límites de confianza unilaterales calculados para μ con σ desconocida son como el lector esperaría, a saber:

$$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}} \qquad \text{y} \qquad x - t_{\alpha}\frac{s}{\sqrt{n}}.$$

Éstos son, respectivamente, los límites superior e inferior del $100(1-\alpha)\%$. Aquí t_{α} es el valor t que tiene una área α a la derecha.

Ejemplo 9.5: El contenido de ácido sulfúrico de 7 contenedores similares es de 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, y 9.6 litros. Calcule un intervalo de confianza del 95% para el contenido promedio de todos los contenedores suponiendo una distribución aproximadamente normal.

***Solución*:** La media muestral y la desviación estándar para los datos dados son

$$\bar{x} = 10.0 \quad \text{y} \quad s = 0.283.$$

Si usamos la tabla A.4, encontramos $t_{0.025} = 2.447$ para $v = 6$ grados de libertad. En consecuencia, el intervalo de confianza del 95% para μ es

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

que se reduce a $9.74 < \mu < 10.26$.

Concepto de intervalo de confianza para una muestra grande

Con frecuencia los estadísticos recomiendan que incluso cuando no sea posible suponer la normalidad, se desconozca σ y $n \geq 30$, σ se puede reemplazar con s para poder utilizar el intervalo de confianza

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

A menudo se hace referencia a esto como un *intervalo de confianza para una muestra grande*. La justificación para esto reside sólo en la presunción de que, con una muestra tan grande como 30 y una distribución de la población no muy sesgada, s estará muy cerca de la σ verdadera y, de esta manera, el teorema del límite central continuará siendo válido. Se debería destacar que esto es sólo una aproximación y que la calidad de los resultados mejora a medida que aumenta el tamaño de la muestra.

Ejemplo 9.6: Se obtienen las calificaciones de matemáticas del Examen de Aptitudes Escolares (SAT, por sus siglas en inglés) de una muestra aleatoria de 500 estudiantes del último año de preparatoria del estado de Texas. Se calculan la media y la desviación estándar muestrales, que son 501 y 112, respectivamente. Calcule un intervalo de confianza del 99% de la calificación promedio de matemáticas en el SAT para los estudiantes del último año de preparatoria del estado de Texas.

***Solución*:** Como el tamaño de la muestra es grande, es razonable utilizar la aproximación normal. Si utilizamos la tabla A.3, encontramos $z_{0.005} = 2.575$. Por lo tanto, un intervalo de confianza del 99% para μ es

$$501 \pm (2.575)\left(\frac{112}{\sqrt{500}}\right) = 501 \pm 12.9,$$

que da como resultado $488.1 < \mu < 513.9$.

9.5 Error estándar de una estimación puntual

Hicimos una distinción muy clara entre los objetivos de las estimaciones puntuales y las estimaciones del intervalo de confianza. Las primeras proporcionan un solo número que se extrae de un conjunto de datos experimentales, y las segundas proporcionan un intervalo razonable para el parámetro, *dados los datos experimentales*; es decir, $100(1-\alpha)\%$ de tales intervalos que se calcula "cubren" el parámetro.

Estos dos métodos de estimación se relacionan entre sí. El elemento en común es la distribución muestral del estimador puntual. Considere, por ejemplo, el estimador $\bar{X}$ de μ cuando se conoce σ. Indicamos antes que una medida de la calidad de un estimador insesgado es su varianza. La varianza de $\bar{X}$ es

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

De esta forma, la desviación estándar de $\bar{X}$ o *error estándar de* $\bar{X}$ es $\sigma/\sqrt{n}$. En términos simples, el error estándar de un estimador es su desviación estándar. Para el caso de $\bar{X}$ el límite de confianza que se calcula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ se escribe como } \bar{x} \pm z_{\alpha/2} \text{ e.e.}(\bar{x}),$$

donde "e.e." es el error estándar. El punto importante es que el ancho del intervalo de confianza de μ depende de la calidad del estimador puntual a través de su error estándar. En el caso en que se desconoce σ y la muestra proviene de una distribución normal, s reemplaza a σ y se incluye el *error estándar estimado* $S/\sqrt{n}$. Por consiguiente, los límites de confianza de μ son:

Límites de confianza para μ cuando se desconoce σ^2

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2} \text{ e.e.}(\bar{x})$$

De nuevo, el intervalo de confianza *no es mejor* (en términos de anchura) *que la calidad de la estimación puntual,* en este caso a través de su error estándar estimado. A menudo el software de computación se refiere a los errores estándar estimados simplemente como "errores estándar".

A medida que avanzamos a intervalos de confianza más complejos, prevalece el concepto de que el ancho de los intervalos de confianza se acorta cuando mejora la calidad de la estimación puntual correspondiente, aunque no siempre es tan sencillo como aquí se ilustra. Se puede argumentar que un intervalo de confianza es tan sólo una ampliación de la estimación puntual para tomar en cuenta la exactitud de dicha estimación.

9.6 Intervalos de predicción

La estimación puntual y la estimación por intervalos de la media que se expusieron en las secciones 9.4 y 9.5 proporcionan buena información del parámetro desconocido μ de una distribución normal, o de una distribución no normal a partir de la cual se toma una muestra grande. Algunas veces, además de la media de la población, el experimentador podría estar interesado en predecir el **valor posible de una observación futura**. Por ejemplo, en el control de calidad el experimentador podría necesitar utilizar los datos observados para predecir una nueva observación. Un proceso de manufactura de una pieza de metal se podría evaluar basándose en si la pieza cumple con las especificaciones de resistencia a la tensión. En ciertas ocasiones un cliente podría estar interesado en comprar una **sola pieza**. En este caso un intervalo de confianza de la resistencia media a la tensión no cubriría la información requerida. El cliente necesitaría una aseveración respecto a la incertidumbre de una **sola observación**. Este tipo de requerimiento se satisface muy bien construyendo un **intervalo de predicción**.

Es muy sencillo obtener un intervalo de predicción para las situaciones que hemos considerado hasta el momento. Suponga que la muestra aleatoria se tomó de una población normal con media μ desconocida y varianza σ^2 conocida. Un estimador puntual natural de una nueva observación es $\bar{X}$. En la sección 8.4 se aprendió que la varianza de $\bar{X}$ es σ^2/n. Sin embargo, para predecir una nueva observación no basta con explicar la variación debida a la estimación de la media, también tendríamos que explicar la **variación de una observación futura**. A partir de la suposición sabemos que la varianza del

error aleatorio en una nueva observación es σ^2. El desarrollo de un intervalo de predicción se representa mejor empezando con una variable aleatoria normal $x_0 - \bar{x}$, donde x_0 es la nueva observación y $\bar{x}$ se toma de la muestra. Como x_0 y $\bar{x}$ son independientes, sabemos que

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + 1/n}}$$

es $n(z; 0, 1)$. Como resultado, si utilizamos el enunciado de probabilidad

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

con el estadístico z anterior, y si colocamos x_0 en el centro del enunciado de probabilidad, tenemos que la probabilidad de que ocurra el siguiente evento es $1 - \alpha$:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}.$$

Como resultado, el intervalo de predicción calculado se formaliza como sigue.

Intervalo de predicción para una observación futura cuando se conoce σ^2

Para una distribución normal de mediciones con media μ desconocida y varianza σ^2 conocida, un **intervalo de predicción** del $100(1 - \alpha)\%$ de una observación futura x_0 es

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Ejemplo 9.7: Debido a la disminución en las tasas de interés el First Citizens Bank recibió muchas solicitudes para hipoteca. Una muestra reciente de 50 créditos hipotecarios dio como resultado un promedio en la cantidad de préstamos de \$257,300. Suponga una desviación estándar de la población de \$25,000. En el caso del siguiente cliente que llena una solicitud de crédito hipotecario calcule un intervalo de predicción del 95% para la cantidad del crédito.

***Solución*:** La predicción puntual de la cantidad del crédito del siguiente cliente es $\bar{x} = \$257{,}300$. El valor z aquí es $z_{0.025} = 1.96$. Por lo tanto, un intervalo de predicción del 95% para la cantidad de un crédito futuro es

$$257,300 - (1.96)(25{,}000)\sqrt{1 + 1/50} < x_0 < 257,300 + (1.96)(25{,}000)\sqrt{1 + 1/50},$$

que produce el intervalo (\$207,812.43, \$306,787.57).

El intervalo de predicción proporciona un buen estimado de la ubicación de una observación futura, el cual es muy diferente del estimado del valor promedio de la muestra. Debe advertirse que la variación de esta predicción es la suma de la variación debida a una estimación de la media y la variación de una sola observación. Sin embargo, como antes, consideramos primero el caso en el que se conoce la varianza. En el caso en que se desconoce la varianza también es importante tratar con el intervalo de predicción de una observación futura. De hecho, en este caso se podría utilizar una distribución t de Student, como se describe en el siguiente resultado. Aquí la distribución normal simplemente se reemplaza con la distribución t.

Intervalo de predicción de una observación futura cuando se desconoce σ^2

Para una distribución normal de mediciones cuando la media μ y la varianza σ^2 se desconocen, un **intervalo de predicción** del $100(1-\alpha)\%$ de una observación futura x_0 es

$$\bar{x} - t_{\alpha/2} s\sqrt{1+1/n} < x_0 < \bar{x} + t_{\alpha/2} s\sqrt{1+1/n},$$

donde $t_{\alpha/2}$ es el valor t con $v = n-1$ grados de libertad, que deja una área de $\alpha/2$ a la derecha.

También se pueden utilizar intervalos de predicción unilaterales. Los límites de predicción superiores se aplican en casos en los que es necesario enfocarse en observaciones futuras grandes. El interés por observaciones pequeñas futuras requiere utilizar límites de predicción más bajos. El límite superior es dado por

$$\bar{x} + t_{\alpha} s\sqrt{1+1/n}$$

y el límite inferior por

$$\bar{x} - t_{\alpha} s\sqrt{1+1/n}.$$

Ejemplo 9.8: Un inspector de alimentos seleccionó aleatoriamente 30 paquetes de carne de res 95% magra. La muestra dio como resultado una media de 96.2% con una desviación estándar muestral de 0.8%. Calcule un intervalo de predicción del 99% para la condición baja en grasa de un paquete nuevo. Suponga normalidad.

***Solución*:** Para $v = 29$ grados de libertad, $t_{0.005} = 2.756$. Por lo tanto, un intervalo de predicción del 99% para una observación nueva x_0 es

$$96.2 - (2.756)(0.8)\sqrt{1+\frac{1}{30}} < x_0 < 96.2 + (2.756)(0.8)\sqrt{1+\frac{1}{30}},$$

que se reduce a (93.96, 98.44).

Uso de límites de predicción para detectar valores extremos

Hasta el momento hemos puesto poca atención al concepto de **valores extremos** u observaciones aberrantes. La mayoría de los investigadores científicos son muy sensibles a la existencia de observaciones de valores extremos, también llamados datos defectuosos o "malos". En el capítulo 12 profundizaremos en el estudio de este concepto. Sin embargo, nos interesa considerarlos aquí porque la detección de los valores extremos está estrechamente relacionada con los intervalos de predicción.

Para nuestros propósitos nos conviene considerar que una observación extrema es una que proviene de una población con una media diferente a la que determina el resto de la muestra de tamaño n que se está estudiando. El intervalo de predicción produce un límite que "cubre" una sola observación futura con probabilidad $1-\alpha$, si ésta proviene de la población de la que se tomó la muestra. Por lo tanto, una metodología para detectar valores extremos implica la regla de que **una observación es un valor extremo si cae fuera del intervalo de predicción calculado sin incluir la observación cuestionable en la muestra**. Como resultado, para el intervalo de predicción del ejemplo 9.8, en el caso de los paquetes de carne, la observación que se obtiene al medir un nuevo paquete y encontrar que su contenido libre de grasa está fuera del intervalo (93.96, 98.44) se podría considerar como un valor extremo.

9.7 Límites de tolerancia

Como vimos en la sección 9.6, el científico o el ingeniero podrían estar menos interesados en estimar parámetros que en obtener información sobre el lugar en el que caería una *observación* o medición individual. Este tipo de situaciones requiere intervalos de predicción. Sin embargo, existe un tercer tipo de intervalo que es útil en muchas aplicaciones. Una vez más, suponga que el interés se centra en torno a la fabricación de la pieza de un componente y que existen especificaciones sobre una dimensión de esa parte. Además, la media de esa dimensión no es tan importante. Sin embargo, a diferencia del escenario de la sección 9.6, se podría estar menos interesado en una sola observación y más en el lugar en el que cae la mayoría de la población. Si las especificaciones del proceso son importantes, el administrador del proceso se interesará en el desempeño a largo plazo, **no en la siguiente observación**. Debemos tratar de determinar los límites que, en cierto sentido probabilístico, "cubren" los valores en la población, es decir, los valores medidos de la dimensión.

Un método para establecer el límite deseado consiste en determinar un intervalo de confianza sobre una *proporción fija* de las mediciones. Esto se comprende mejor visualizando una situación en la que se realiza un muestreo aleatorio de una distribución normal con media conocida μ y varianza σ^2. Evidentemente, un límite que cubre el 95% central de la población de observaciones es

$$\mu \pm 1.96\sigma.$$

A esto se le llama **intervalo de tolerancia** y, en realidad, su cobertura del 95% de las observaciones medidas es exacta. Sin embargo, en la práctica rara vez se conocen μ y σ; por consiguiente, el usuario debe aplicar

$$\bar{x} \pm ks.$$

Ahora bien, el intervalo es, desde luego, una variable aleatoria, por lo tanto, la *cobertura* de una proporción de la población por el intervalo no es exacta. Como resultado, se debe usar un intervalo de confianza del $100(1-\gamma)\%$, ya que no se puede esperar que $\bar{x} \pm ks$ cubra cualquier proporción específica todo el tiempo. Lo anterior nos lleva a la siguiente definición.

Límites de tolerancia

Para una distribución normal de mediciones en la que se desconoce la media μ y la desviación estándar σ, los **límites de tolerancia** son dados por $\bar{x} \pm ks$, donde k se determina de tal manera que se pueda estar seguro, con un $100(1-\gamma)\%$ de confianza, de que los límites dados contienen al menos la proporción $1-\alpha$ de las mediciones.

La tabla A.7 ofrece valores de k para $1-\alpha = 0.90, 0.95, 0.99$; $\gamma = 0.05, 0.01$; y para valores seleccionados de n de 2 a 300.

Ejemplo 9.9: Considere el ejemplo 9.8. Con la información dada calcule un intervalo de tolerancia que proporcione límites bilaterales del 95% sobre el 90% de la distribución de paquetes de carne 95% magra. Suponga que los datos provienen de una distribución aproximadamente normal.

***Solución*:** Del ejemplo 9.8, recuerde que $n = 30$, que la media muestral es de 96.2% y que la desviación estándar muestral es de 0.8%. De la tabla A.7, $k = 2.14$. Si utilizamos

$$\bar{x} \pm ks = 96.2 \pm (2.14)(0.8),$$

encontramos que los límites inferior y superior son de 94.5 y de 97.9.

Tenemos 95% de confianza en que el rango anterior cubre el 90% central de la distribución de paquetes de carne de res 95% magra.

Diferencia entre intervalos de confianza, intervalos de predicción e intervalos de tolerancia

Es importante resaltar la diferencia entre los tres tipos de intervalos que se estudiaron e ilustraron en las secciones anteriores. Los cálculos son sencillos, pero la interpretación podría resultar confusa. En aplicaciones de la vida real tales intervalos no son intercambiables, ya que sus interpretaciones son muy diferentes.

En el caso de los intervalos de confianza sólo se pone atención en la **media de la población**. Por ejemplo, el ejercicio 9.13 de la página 283 se refiere a un proceso de ingeniería que produce alfileres para costura. Se establece una especificación sobre la dureza de Rockwell por debajo de la cual el cliente no aceptará ningún alfiler. En este caso un parámetro de la población debe tener poca relevancia. Es importante que el ingeniero sepa en dónde *van a estar la mayoría de los valores de la dureza de Rockwell*. Por consiguiente, se deberían utilizar los límites de tolerancia. Seguramente, al administrador le agradará saber que los límites de tolerancia en cualquier producto del proceso son más rigurosos que las especificaciones para el propio proceso.

Es verdad que la interpretación del límite de tolerancia se relaciona hasta cierto punto con el intervalo de confianza. El intervalo de tolerancia del $100(1-\alpha)\%$ sobre, digamos, la proporción 0.95, se podría considerar como un intervalo de confianza **sobre el 95% intermedio** de la distribución normal correspondiente. Los límites de tolerancia unilaterales también son relevantes. En el caso del problema de dureza de Rockwell se desearía tener un límite inferior de la forma $\bar{x} - ks$, tal que se tenga un 99% de confianza en que al menos 99% de los valores de la dureza de Rockwell excederán al valor calculado.

Los intervalos de predicción se pueden aplicar cuando es importante determinar un límite para un **solo valor**. Aquí la media no es la cuestión, ni tampoco la ubicación de la mayoría de la población, lo que se requiere, más bien, es la ubicación de una sola nueva observación.

Estudio de caso 9.1: **Calidad de una máquina.** Una máquina produce piezas de metal que tienen forma cilíndrica. Se toma una muestra de tales piezas y se encuentra que los diámetros son 1.01, 0.97. 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 y 1.03 centímetros. Utilice estos datos para calcular tres tipos de intervalos y hacer interpretaciones que ilustren las diferencias entre ellos en el contexto del sistema. Para todos los cálculos suponga una distribución aproximadamente normal. La media muestral y la desviación estándar para los datos dados son $\bar{x} = 1.0056$ y $s = 0.0246$.

a) Calcule un intervalo de confianza del 99% sobre la media del diámetro.

b) Calcule un intervalo de predicción del 99% sobre el diámetro medido de una sola pieza de metal tomada de la máquina.

c) Calcule los límites de tolerancia del 99% que contengan 95% de las piezas de metal producidas por esta máquina.

***Solución*:** *a)* El intervalo de confianza del 99% para la media del diámetro está dado por

$$\bar{x} \pm t_{0.005}\, s/\sqrt{n} = 1.0056 \pm (3.355)(0.0246/3) = 1.0056 \pm 0.0275.$$

Por lo tanto, los límites de confianza del 99% son 0.9781 y 1.0331.

b) El intervalo de predicción del 99% para una futura observación está dado por

$$\bar{x} \pm t_{0.005}\, s\sqrt{1 + 1/n} = 1.0056 \pm (3.355)(0.0246)\sqrt{1 + 1/9},$$

donde los límites son 0.9186 y 1.0926.

c) De la tabla A.7, para $n = 9$, $1 - \gamma = 0.99$, y $1 - \alpha = 0.95$, obtenemos $k = 4.550$ para los límites bilaterales. Por lo tanto, los límites de tolerancia del 99% son dados por

$$\bar{x} + ks = 1.0056 \pm (4.550)(0.0246),$$

donde los límites son 0.8937 y 1.1175. Tenemos un 99% de confianza en que el intervalo de tolerancia de 0.8937 a 1.1175 contendrá el 95% central de la distribución de diámetros producidos.

Este estudio de caso ilustra que los tres tipos de límites pueden conducir a resultados muy diferentes, aunque todos son límites del 99%. En el caso del intervalo de confianza sobre la media, el 99% de estos intervalos cubre la media del diámetro de la población. Por lo tanto, decimos que tenemos un 99% de confianza en que la media del diámetro producido por el proceso se encuentra entre 0.9781 y 1.0331 centímetros. Se hace hincapié en la media y se pone poco interés en una sola lectura o en la naturaleza general de la distribución de diámetros en la población. En lo que se refiere a los límites de predicción, los límites 0.9186 y 1.0926 se basan en la distribución de una sola pieza "nueva" de metal tomada del proceso, y nuevamente el 99% de estos límites cubren el diámetro de una nueva pieza medida. Por otro lado, como se sugirió en la sección anterior, los límites de tolerancia le dan al ingeniero una idea de en qué parte de la población se localiza la "mayoría", digamos el 95% central, de los diámetros de las piezas medidas. Los límites de tolerancia del 99%, 0.8937 y 1.1175 difieren mucho de los otros dos límites. Si esos límites le parecen demasiado anchos al ingeniero, esto se reflejará de forma negativa en la calidad del proceso. Por otro lado, si los límites representan un resultado deseable, el ingeniero podría concluir que la mayoría (95% en este caso) de los diámetros se encuentran dentro de un rango adecuado. De nuevo, se podría hacer una interpretación del intervalo de confianza, a saber, el 99% de esos límites calculados cubrirán el 95% intermedio de la población de diámetros. ■

Ejercicios

9.1 Un investigador de la UCLA afirma que la esperanza de vida de los ratones se puede extender hasta en 25% cuando se reduce aproximadamente 40% de las calorías de su dieta desde el momento en que son destetados. La dieta restringida se enriquece hasta niveles normales con vitaminas y proteínas. Si se supone que a partir de estudios previos se sabe que $\sigma = 5.8$ meses, ¿cuántos ratones se deberían incluir en la muestra para tener un 99% de confianza en que la vida media esperada de la muestra estará dentro de 2 meses a partir de la media de la población para todos los ratones sujetos a la dieta reducida?

9.2 Una empresa de material eléctrico fabrica bombillas que tienen una duración distribuida de forma aproximadamente normal, con una desviación estándar de 40 horas. Si una muestra de 30 bombillas tiene una duración promedio de 780 horas, calcule un intervalo de confianza del 96% para la media de la población de todas las bombillas producidas por esta empresa.

9.3 Muchos pacientes con problemas del corazón tienen un marcapasos para controlar su ritmo cardiaco. El marcapasos tiene montado un módulo conector de plástico en la parte superior. Suponga una desviación estándar de 0.0015 pulgadas y una distribución aproximadamente normal, y con base en esto calcule un intervalo de confianza del 95% para la media de la profundidad de todos los módulos conectores fabricados por cierta empresa. Una muestra aleatoria de 75 módulos tiene una profundidad promedio de 0.310 pulgadas.

9.4 Las estaturas de una muestra aleatoria de 50 estudiantes universitarios tienen una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros.

a) Construya un intervalo de confianza del 98% para la estatura media de todos los estudiantes universitarios.

b) ¿Qué podemos afirmar con una confianza del 98% acerca del posible tamaño de nuestro error, si estimamos que la estatura media de todos los estudiantes universitarios es de 174.5 centímetros?

9.5 Una muestra aleatoria de 100 propietarios de automóviles del estado de Virginia revela que éstos conducen su automóvil, en promedio, 23,500 kilómetros por año, con una desviación estándar de 3900 kilómetros. Suponga que la distribución de las mediciones es aproximadamente normal.

a) Construya un intervalo de confianza del 99% para el número promedio de kilómetros que un propietario de un automóvil conduce anualmente en Virginia.

b) ¿Qué podemos afirmar con un 99% de confianza acerca del posible tamaño del error, si estimamos que los propietarios de automóviles de Virginia conducen un promedio de 23,500 kilómetros por año?

9.6 ¿Qué tan grande debe ser la muestra en el ejercicio 9.2 si deseamos tener un 96% de confianza en que nuestra media muestral estará dentro de 10 horas a partir de la media verdadera?

9.7 ¿De qué tamaño debe ser la muestra en el ejercicio 9.3 si deseamos tener un 95% de confianza en que nuestra media muestral estará dentro de un 0.0005 de pulgada de la media verdadera?

9.8 Un experto en eficiencia desea determinar el tiempo promedio que toma perforar tres hoyos en cierta placa metálica. ¿De qué tamaño debe ser una muestra para tener un 95% de confianza en que esta media muestral estará dentro de 15 segundos de la media verdadera? Suponga que por estudios previos se sabe que $\sigma = 40$ segundos.

9.9 Según estudios realizados por el doctor W. H. Bowen, del Instituto Nacional de Salud, y por el doctor J. Yudben, profesor de nutrición y dietética de la Universidad de Londres, el consumo regular de cereales preendulzados contribuye al deterioro de los dientes, a las enfermedades cardiacas y a otras enfermedades degenerativas. En una muestra aleatoria de 20 porciones sencillas similares del cereal Alpha-Bits, el contenido promedio de azúcar era de 11.3 gramos con una desviación estándar de 2.45 gramos. Suponga que el contenido de azúcar está distribuido normalmente y con base en esto construya un intervalo de confianza de 95% para el contenido medio de azúcar de porciones sencillas de Alpha-Bits.

9.10 Las integrantes de una muestra aleatoria de 12 graduadas de cierta escuela para secretarias teclearon un promedio de 79.3 palabras por minuto, con una desviación estándar de 7.8 palabras por minuto. Suponga una distribución normal para el número de palabras que teclean por minuto y con base en esto calcule un intervalo de confianza del 95% para el número promedio de palabras que teclean todas las graduadas de esta escuela.

9.11 Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de las piezas y los diámetros son 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 y 1.03 centímetros. Calcule un intervalo de confianza del 99% para la media del diámetro de las piezas que se manufacturan con esta máquina. Suponga una distribución aproximadamente normal.

9.12 Una muestra aleatoria de 10 barras energéticas de chocolate de cierta marca tiene, en promedio, 230 calorías por barra y una desviación estándar de 15 calorías. Construya un intervalo de confianza del 99% para el contenido medio verdadero de calorías de esta marca de barras energéticas de chocolate. Suponga que la distribución del contenido calórico es aproximadamente normal.

9.13 En un estudio para determinar la dureza de Rockwell en la cabeza de alfileres para costura se toma una muestra aleatoria de 12. Se toman mediciones de la dureza de Rockwell para cada una de las 12 cabezas y se obtiene un valor promedio de 48.50, con una desviación estándar muestral de 1.5. Suponga que las mediciones se distribuyen de forma normal y con base en esto construya un intervalo de confianza de 90% para la dureza media de Rockwell.

9.14 Se registran las siguientes mediciones del tiempo de secado, en horas, de cierta marca de pintura vinílica:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones representan una muestra aleatoria de una población normal y con base en esto calcule el intervalo de predicción del 95% para el tiempo de secado de la siguiente prueba de pintura.

9.15 Remítase al ejercicio 9.5 y construya un intervalo de predicción del 99% para los kilómetros que viaja anualmente el propietario de un automóvil en Virginia.

9.16 Considere el ejercicio 9.10 y calcule el intervalo de predicción del 95% para el siguiente número observado de palabras por minuto tecleadas por una graduada de la escuela de secretarias.

9.17 Considere el ejercicio 9.9 y calcule un intervalo de predicción del 95% para el contenido de azúcar de la siguiente porción de cereal Alpha-Bits.

9.18 Remítase al ejercicio 9.13 y construya un intervalo de tolerancia del 95% que contenga el 90% de las mediciones.

9.19 Una muestra aleatoria de 25 tabletas de aspirina con antiácido contiene, en promedio, 325.05 mg de aspirina en cada tableta, con una desviación estándar de 0.5 mg. Calcule los límites de tolerancia del 95% que contendrán 90% del contenido de aspirina para esta marca. Suponga que el contenido de aspirina se distribuye normalmente.

9.20 Considere la situación del ejercicio 9.11. Aunque la estimación de la media del diámetro es importante, no es ni con mucho tan importante como intentar determinar la ubicación de la mayoría de la distribución de los diámetros. Calcule los límites de tolerancia del 95% que contengan el 95% de los diámetros.

9.21 En un estudio realizado por el Departamento de Zoología del Virginia Tech con el fin de conocer la cantidad de ortofósforo en el río, se recolectaron 15 "muestras" de agua en una determinada estación ubicada en el río James. La concentración del químico se midió en miligramos por litro. Suponga que la media en la estación de muestreo no es tan importante como la distribución de las concentraciones del químico en los extremos superiores. El interés se centra en saber si las concentraciones en estos extremos son demasiado elevadas. Las lecturas de las 15 muestras de agua proporcionaron una media muestral de 3.84 miligramos por litro y una desviación estándar muestral de 3.07 miligramos por litro. Suponga que las lecturas son una muestra aleatoria de una distribución normal. Calcule un intervalo de predicción (límite de predicción superior del 95%) y un límite de tolerancia (un límite de tolerancia superior del 95% que excede al 95% de la población de valores). Interprete ambos límites, es decir, especifique qué indica cada uno acerca de los extremos superiores de la distribución de ortofósforo en la estación de muestreo.

9.22 Se están estudiando las propiedades de resistencia a la tensión de un determinado tipo de hilo. Con ese fin se prueban 50 piezas en condiciones similares y los resultados que se obtienen revelan una resistencia a la tensión promedio de 78.3 kilogramos y una desviación estándar de 5.6 kilogramos. Suponga que la resistencia a la tensión tiene una distribución normal y con base en esto calcule un límite de predicción inferior al 95% de un solo valor observado de resistencia a la tensión. Además, determine un límite inferior de tolerancia del 95% que sea excedido por el 99% de los valores de resistencia a la tensión.

9.23 Remítase al ejercicio 9.22. ¿Por qué las 1/2 cantidades solicitadas en el ejercicio parecen ser más importantes para el fabricante del hilo que, por ejemplo, un intervalo de confianza en la resistencia media a la tensión?

9.24 Remítase una vez más al ejercicio 9.22. Suponga que un comprador del hilo especifica que éste debe tener una resistencia a la tensión de por lo menos 62 kilogramos. El fabricante estará satisfecho si la cantidad de piezas producidas que no cumplen la especificación no excede al 5%. ¿Hay alguna razón para preocuparse? Esta vez utilice un límite de tolerancia unilateral del 99% que sea excedido por el 95% de los valores de resistencia a la tensión.

9.25 Considere las mediciones del tiempo de secado del ejercicio 9.14. Suponga que las 15 observaciones en el conjunto de datos también incluyen un decimosexto valor de 6.9 horas. En el contexto de las 15 observaciones originales, ¿el valor decimosexto es un valor extremo? Muestre el procedimiento.

9.26 Considere los datos del ejercicio 9.13. Suponga que el fabricante de los alfileres insiste en que la dureza de Rockwell del producto es menor o igual que 44.0 sólo un 5% de las veces. ¿Cuál es su reacción? Utilice un cálculo de un límite de tolerancia como la base de su veredicto.

9.27 Considere la situación del estudio de caso 9.1 de la página 281, con una muestra más grande de piezas metálicas. Los diámetros son los siguientes: 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 1.01, 1.03, 0.99, 1.00, 1.00, 0.99, 0.98, 1.01, 1.02, 0.99 centímetros. Nuevamente puede suponer una distribución normal. Haga lo siguiente y compare sus resultados con los del estudio de caso. Analice en qué difieren y por qué.

a) Calcule un intervalo de confianza del 99% de la media del diámetro.
b) Calcule un intervalo de predicción del 99% en la medición del siguiente diámetro.
c) Calcule un intervalo de tolerancia del 99% para la cobertura del 95% central de la distribución de diámetros.

9.28 En la sección 9.3 destacamos el concepto del "estimador más eficaz" comparando la varianza de dos estimadores insesgados $\hat{\Theta}_1$ y $\hat{\Theta}_2$. Sin embargo, esto no toma en cuenta el sesgo en el caso en que uno o ambos estimadores no son sesgados. Considere la cantidad

$$EME = E(\hat{\Theta} - \theta),$$

donde EME denota el **error cuadrático medio**. El error cuadrático medio a menudo se utiliza para comparar dos estimadores $\hat{\Theta}_1$ y $\hat{\Theta}_2$ de θ, cuando uno o ambos no son sesgados porque i) es intuitivamente razonable y ii) se toma en cuenta para el sesgo. Demuestre que el EME se puede escribir como

$$EME = E[\hat{\Theta} - E(\hat{\Theta})]^2 + [E(\hat{\Theta} - \theta)]^2$$
$$= \text{Var}(\hat{\Theta}) + [\text{sesgo}(\hat{\Theta})]^2.$$

9.29 Definamos $S'^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$. Demuestre que

$$E(S'^2) = [(n-1)/n]\sigma^2,$$

y, en consecuencia, que S'^2 es un estimador sesgado para σ^2.

9.30 Considere S'^2, el estimador de σ^2, del ejercicio 9.29. Con frecuencia los analistas utilizan S'^2 en lugar de dividir $\sum_{i=1}^{n}(X_i - \bar{X})^2$ entre $n-1$, los grados de libertad en la muestra.

a) ¿Cuál es el sesgo de S'^2?
b) Demuestre que el sesgo de S'^2 se aproxima a cero a medida que $n \to \infty$.

9.31 Si X es una variable aleatoria binomial, demuestre que

a) $\hat{P} = X/n$ es un estimador insesgado de p;
b) $P' = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$ es un estimador sesgado de p.

9.32 Demuestre que el estimador P' del ejercicio 9.31*b*) se vuelve no sesgado a medida que $n \to \infty$.

9.33 Compare S^2 y S'^2 (véase el ejercicio 9.29), los dos estimadores de σ^2, para determinar cuál es más eficaz. Suponga que estos estimadores se obtienen usando $X_1, X_2, \ldots, X_n$, las variables aleatorias independientes de $n(x; \mu, \sigma)$. ¿Cuál es el estimador más eficaz si se considera sólo la varianza de los estimadores? [*Sugerencia*: Utilice el teorema 8.4 y el hecho de que la varianza de χ_v^2 es $2v$, de la sección 6.7.]

9.34 Considere el ejercicio 9.33. Utilice el *EME* que se estudió en el ejercicio 9.28 para determinar qué estimador es más eficaz. Escriba

$$\frac{EME\ (S^2)}{EME\ (S'^2)}.$$

9.8 Dos muestras: estimación de la diferencia entre dos medias

Si tenemos dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2, respectivamente, el estadístico que da un estimador puntual de la diferencia entre μ_1 y μ_2 es $\bar{X}_1 - \bar{X}_2$. Por lo tanto, para obtener una estimación puntual de $\mu_1 - \mu_2$, se seleccionan dos muestras aleatorias independientes, una de cada población, de tamaños n_1 y n_2, y se calcula $\bar{x}_1 - \bar{x}_2$, la diferencia de las medias muestrales. Evidentemente, debemos considerar la distribución muestral de $\bar{X}_1 - \bar{X}_2$.

De acuerdo con el teorema 8.3, podemos esperar que la distribución muestral de $\bar{X}_1 - \bar{X}_2$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ y desviación estándar $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Por lo tanto, podemos asegurar, con una probabilidad de $1 - \alpha$, que la variable normal estándar

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

caerá entre $-z_{\alpha/2}$ y $z_{\alpha/2}$. Si nos remitimos una vez más a la figura 9.2, escribimos

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Al sustituir para Z, establecemos de manera equivalente que

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

que conduce al siguiente intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$.

Intervalo de confianza para $\mu_1 - \mu_2$ cuando se conocen σ_1^2 y σ_2^2

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños n_1 y n_2, de poblaciones que tienen varianzas conocidas σ_1^2 y σ_2^2, respectivamente, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

El grado de confianza es exacto cuando las muestras se seleccionan de poblaciones normales. Para poblaciones no normales el teorema del límite central permite una buena aproximación para muestras de tamaño razonable.

Las condiciones experimentales y la unidad experimental

Para el caso en que se necesita estimar un intervalo de confianza sobre la diferencia entre dos medias se requiere considerar las condiciones experimentales durante el proceso de recolección de datos. Se supone que tenemos dos muestras aleatorias independientes de distribuciones con medias μ_1 y μ_2, respectivamente. Es importante que las condiciones experimentales se parezcan al ideal descrito por las suposiciones tanto como sea posible. Con mucha frecuencia el experimentador debería planear la estrategia del experimento de acuerdo con esto. Para casi cualquier estudio de este tipo existe una *unidad experimental*, que es la parte del experimento que produce el error experimental y genera la varianza de la población que denominamos σ^2. En un estudio farmacológico la unidad experimental es el paciente o el sujeto. En un experimento de agricultura puede ser una superficie de tierra. En un experimento químico puede ser una cantidad de materias primas. Es importante que las diferencias entre tales unidades tengan un impacto mínimo sobre los resultados. El experimentador tendrá un grado de seguridad de que las unidades experimentales no sesgarán los resultados si las condiciones que definen a las dos poblaciones se *asignan al azar a* las unidades experimentales. En los siguientes capítulos acerca de la prueba de hipótesis nos volveremos a concentrar en la aleatorización.

Ejemplo 9.10: Se llevó a cabo un experimento donde se compararon dos tipos de motores, el *A* y el *B*. Se midió el rendimiento de combustible en millas por galón. Se realizaron 50 experimentos con el motor tipo *A* y 75 con el motor tipo *B*. La gasolina utilizada y las demás condiciones se mantuvieron constantes. El rendimiento promedio de gasolina para el motor *A* fue de 36 millas por galón y el promedio para el motor *B* fue de 42 millas por galón. Calcule un intervalo de confianza del 96% sobre $\mu_B - \mu_A$, donde μ_A y μ_B corresponden a la media de la población del rendimiento de millas por galón para los motores *A* y *B*, respectivamente. Suponga que las desviaciones estándar de la población son 6 y 8 para los motores *A* y *B*, respectivamente.

***Solución*:** La estimación puntual de $\mu_B - \mu_A$ es $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$. Si usamos $\alpha = 0.04$, obtenemos $z_{0.02} = 2.05$ de la tabla A.3. Por lo tanto, sustituyendo en la fórmula anterior, el intervalo de confianza del 96% es

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}},$$

o simplemente $3.43 < \mu_B - \mu_A < 8.57$. ∎

Este procedimiento para estimar la diferencia entre dos medias se aplica si se conocen σ_1^2 y σ_2^2. Si las varianzas no se conocen y las dos distribuciones implicadas son aproximadamente normales, la distribución *t* resulta implicada como en el caso de una sola muestra. Si no se está dispuesto a suponer normalidad, muestras grandes (digamos mayores que 30) permitirán usar s_1 y s_2 en lugar de σ_1 y σ_2, respectivamente, con el fundamento de que $s_1 \approx \sigma_1$ y $s_2 \approx \sigma_2$. De nuevo, por supuesto, el intervalo de confianza es aproximado.

Varianzas desconocidas pero iguales

Considere el caso donde se desconocen σ_1^2 y σ_2^2. Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ obtenemos una variable normal estándar de la forma

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}.$$

De acuerdo con el teorema 8.4, las dos variables aleatorias

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \quad \text{y} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

tienen distribuciones chi cuadrada con $n_1 - 1$ y $n_2 - 1$ grados de libertad, respectivamente. Además, son variables chi cuadrada independientes, ya que las muestras aleatorias se seleccionaron de forma independiente. En consecuencia, su suma

$$V = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n_1 + n_2 - 2$ grados de libertad.

Como se puede demostrar que las expresiones anteriores para Z y V son independientes, del teorema 8.5 se sigue que el estadístico

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \Bigg/ \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2(n_1 + n_2 - 2)}}$$

tiene la distribución t con $v = n_1 + n_2 - 2$ grados de libertad.

Se puede obtener una estimación puntual de la varianza común desconocida σ^2 agrupando las varianzas muestrales. Si representamos con S_p^2 al estimador agrupado, obtenemos lo siguiente,

Estimado agrupado de la varianza

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Al sustituir S_p^2 en el estadístico T, obtenemos la forma menos engorrosa:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}}.$$

Si usamos el estadístico T, tenemos

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor t con $n_1 + n_2 - 2$ grados de libertad, por arriba del cual encontramos una área de $\alpha/2$. Al sustituir por T en la desigualdad, escribimos

$$P\left[-t_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}} < t_{\alpha/2}\right] = 1 - \alpha.$$

Después de realizar las manipulaciones matemáticas de costumbre, se calculan la diferencia de las medias muestrales $\bar{x}_1 - \bar{x}_2$ y la varianza agrupada, y se obtiene el siguiente intervalo de confianza del $100(1 - \alpha)\%$ para $\mu_1 - \mu_2$.

Se observa con facilidad que el valor de s_p^2 es un promedio ponderado de las dos varianzas muestrales s_1^2 y s_2^2, donde los pesos son los grados de libertad.

Intervalo de confianza para $\mu_1 - \mu_2$, $\sigma_1^2 = \sigma_2^2$ cuando se desconocen ambas varianzas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes con tamaños n_1 y n_2, respectivamente, tomadas de poblaciones más o menos normales con varianzas iguales pero desconocidas, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

donde s_p es la estimación agrupada de la desviación estándar de la población y $t_{\alpha/2}$ es el valor t con $v = n_1 + n_2 - 2$ grados de libertad, que deja una área de $\alpha/2$ a la derecha.

Ejemplo 9.11: En el artículo "Estructura comunitaria de los macroinvertebrados como un indicador de la contaminación de minas ácidas", publicado en el *Journal of Environmental Pollution*, se informa sobre una investigación realizada en Cane Creek, Alabama, para determinar la relación entre parámetros fisioquímicos seleccionados y diversas mediciones de la estructura de la comunidad de macroinvertebrados. Una faceta de la investigación consistió en evaluar la efectividad de un índice numérico de la diversidad de especies para indicar la degradación del agua debida al desagüe ácido de una mina. Conceptualmente, un índice elevado de la diversidad de especies macroinvertebradas debería indicar un sistema acuático no contaminado; mientras que un índice bajo de esta diversidad indicaría un sistema acuático contaminado.

Se eligieron 2 estaciones de muestreo independientes para este estudio: una que se localiza corriente abajo del punto de descarga ácida de la mina y la otra ubicada corriente arriba. Para 12 muestras mensuales reunidas en la estación corriente abajo el índice de diversidad de especies tuvo un valor medio de $\bar{x}_1 = 3.11$ y una desviación estándar de $s_1 = 0.771$; mientras que 10 muestras reunidas mensualmente en la estación corriente arriba tuvieron un valor medio del índice $\bar{x}_2 = 2.04$ y una desviación estándar de $s_2 = 0.448$. Calculemos un intervalo de confianza del 90% para la diferencia entre las medias de la población de los dos sitios, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal y que tienen varianzas iguales.

***Solución*:** Representemos con μ_1 y μ_2 las medias de la población para los índices de diversidad de especies en las estaciones corriente abajo y corriente arriba, respectivamente. Deseamos encontrar un intervalo de confianza del 90% para $\mu_1 - \mu_2$. La estimación puntual de $\mu_1 - \mu_2$ es

$$\bar{x}_1 - \bar{x}_2 = 3.11 - 2.04 = 1.07$$

El estimado agrupado, s_p^2, de la varianza común, σ^2, es

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)(0.771^2) + (9)(0.448^2)}{12 + 10 - 2} = 0.417.$$

Al sacar la raíz cuadrada obtenemos $s_p = 0.646$. Si usamos $\alpha = 0.1$, encontramos en la tabla A.4 que $t_{0.05} = 1.725$ para $v = n_1 + n_2 - 2 = 20$ grados de libertad. Por lo tanto, el intervalo de confianza del 90% para $\mu_1 - \mu_2$ es

$$1.07 - (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}},$$

que se simplifica a $0.593 < \mu_1 - \mu_2 < 1.547$. ∎

Interpretación del intervalo de confianza

Para el caso de un solo parámetro el intervalo de confianza simplemente produce límites de error del parámetro. Los valores contenidos en el intervalo se deberían ver como valores razonables, dados los datos experimentales. En el caso de una diferencia entre dos medias, la interpretación se puede extender a una comparación de las dos medias. Por ejemplo, si tenemos gran confianza en que una diferencia $\mu_1 - \mu_2$ es positiva, sin duda inferiremos que $\mu_1 > \mu_2$ con poco riesgo de incurrir en un error. Así, en el ejemplo 9.11 tenemos un 90% de confianza en que el intervalo de 0.593 a 1.547 contiene la diferencia de las medias de la población para valores del índice de diversidad de especies en las dos estaciones. El hecho de que ambos límites de confianza sean positivos indica que, en promedio, el índice para la estación que se localiza corriente abajo del punto de descarga es mayor que el índice para la estación que se localiza corriente arriba.

Muestras de tamaños iguales

El procedimiento para construir intervalos de confianza para $\mu_1 - \mu_2$ cuando $\sigma_1 = \sigma_2 = \sigma$ pero ésta se desconoce, requiere suponer que las poblaciones son normales. Desviaciones ligeras de la suposición de varianzas iguales o de normalidad no alteran seriamente el grado de confianza en nuestro intervalo. (En el capítulo 10 se estudia un procedimiento para probar la igualdad de dos varianzas poblacionales desconocidas con base en la información que proporcionan las varianzas muestrales). Si las varianzas de la población son considerablemente diferentes, aún obtenemos resultados razonables cuando las poblaciones son normales, siempre y cuando $n_1 = n_2$. Por lo tanto, al planear un experimento se debería hacer un esfuerzo por igualar el tamaño de las muestras.

Varianzas desconocidas y distintas

Consideremos ahora el problema de calcular el estimado de un intervalo de $\mu_1 - \mu_2$ cuando no es probable que las varianzas de la población desconocidas sean iguales. El estadístico que se utiliza con mayor frecuencia en este caso es

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}},$$

que tiene aproximadamente una distribución t con v grados de libertad, donde

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}.$$

Como v rara vez es un entero, lo *redondeamos* al número entero menor más cercano. El estimado anterior de los grados de libertad se denomina aproximación de Satterthwaite (Satterthwaite, 1946, en la bibliografía).

Con el estadístico T', escribimos

$$P(-t_{\alpha/2} < T' < t_{\alpha/2}) \approx 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor de la distribución t con v grados de libertad por arriba del cual encontramos una área de $\alpha/2$. Al sustituir para T' en la desigualdad y seguir los mismos pasos que antes, establecemos el resultado final.

Intervalo de confianza para $\mu_1 - \mu_2$, $\sigma_1^2 \neq \sigma_2^2$ y ambas varianzas se desconocen

Si $\bar{x}_1$ y s_1^2 y $\bar{x}_2$ y s_2^2 son las medias y varianzas de muestras aleatorias independientes de tamaños n_1 y n_2, respectivamente, tomadas de poblaciones aproximadamente normales con varianzas desconocidas y diferentes, un intervalo de confianza aproximado del $100(1 - \alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

donde $t_{\alpha/2}$ es el valor t con

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

grados de libertad, que deja una área de $\alpha/2$ a la derecha.

Observe que la expresión para el valor v anterior incluye variables aleatorias y, por consiguiente, v es un *estimado* de los grados de libertad. En las aplicaciones este estimado no será un número entero, de manera que el analista lo debe redondear al entero menor más cercano para lograr la confianza que se busca.

Antes de ilustrar el intervalo de confianza anterior con un ejemplo deberíamos señalar que todos los intervalos de confianza para $\mu_1 - \mu_2$ tienen la misma forma general, como los de una sola media; a saber, se pueden escribir como

$$\text{estimación puntual} \pm t_{\alpha/2}\,\widehat{\text{e.e.}}(\text{estimación puntual})$$

o

$$\text{estimación puntual} \pm z_{\alpha/2}\,\text{e.e.}(\text{estimación puntual}).$$

Por ejemplo, en el caso donde $\sigma_1 = \sigma_2 = \sigma$, el error estándar estimado de $\bar{x}_1 - \bar{x}_2$ es $s_p\sqrt{1/n_1 + 1/n_2}$. Para el caso donde $\sigma_1^2 \neq \sigma_2^2$,

$$\widehat{\text{e.e.}}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Ejemplo 9.12: El Departamento de zoología de Virginia Tech llevó a cabo un estudio para estimar la diferencia en la cantidad de ortofósforo químico medido en dos estaciones diferentes del río James. El ortofósforo se mide en miligramos por litro. Se reunieron 15 muestras de la estación 1 y 12 muestras de la estación 2. Las 15 muestras de la estación 1 tuvieron un contenido promedio de ortofósforo de 3.84 miligramos por litro y una desviación estándar de 3.07 miligramos por litro; en tanto que las 12 muestras de la estación 2 tuvieron un contenido promedio de 1.49 miligramos por litro y una desviación estándar de 0.80 miligramos por litro. Calcule un intervalo de confianza de 95% para la diferencia en el contenido promedio verdadero de ortofósforo en estas dos estaciones. Suponga que las observaciones provienen de poblaciones normales con varianzas diferentes.

***Solución*:** Para la estación 1 tenemos $\bar{x}_1 = 3.84$, $s_1 = 3.07$ y $n_1 = 15$. Para la estación 2, $\bar{x}_2 = 1.49$, $s_2 = 0.80$ y $n_2 = 12$. Queremos obtener un intervalo de confianza del 95% para $\mu_1 - \mu_2$.

Como se suponen varianzas de la población diferentes, sólo podemos calcular un intervalo de confianza aproximado del 95% basado en la distribución t con v grados de libertad, donde

$$v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16.$$

Nuestra estimación puntual de $\mu_1 - \mu_2$ es

$$\bar{x}_1 - \bar{x}_2 = 3.84 - 1.49 = 2.35.$$

Si usamos $\alpha = 0.05$, en la tabla A.4 encontramos que $t_{0.025} = 2.120$ para $v = 16$ grados de libertad. Por lo tanto, el intervalo de confianza del 95% para $\mu_1 - \mu_2$ es

$$2.35 - 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < 2.35 + 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}},$$

que se simplifica a $0.60 < \mu_1 - \mu_2 < 4.10$. En consecuencia, tenemos un 95% de confianza en que el intervalo de 0.60 a 4.10 miligramos por litro contiene la diferencia del promedio verdadero del ortofósforo que contienen estos dos lugares. ∎

Cuando se desconocen dos varianzas de la población, la suposición de varianzas iguales o diferentes podría ser precaria. En la sección 10.10 se presentará un procedimiento que ayudará a distinguir entre las situaciones con la misma varianza y con varianza diferente.

9.9 Observaciones pareadas

Ahora estudiaremos los procedimientos de estimación para la diferencia de dos medias cuando las muestras no son independientes y las varianzas de las dos poblaciones no son necesariamente iguales. La situación que se considera aquí tiene que ver con una condición experimental muy especial, a saber, *las observaciones pareadas*. A diferencia de la situación que se describió antes, las condiciones de las dos poblaciones no se asignan de forma aleatoria a las unidades experimentales. Más bien, cada unidad experimental homogénea recibe ambas condiciones de la población; como resultado, cada unidad experimental tiene un par de observaciones, una para cada población. Por ejemplo, si realizamos una prueba de una nueva dieta con 15 individuos, los pesos antes y después de seguir la dieta conforman la información de las dos muestras. Las dos poblaciones son "antes" y "después", y la unidad experimental es el individuo. Evidentemente, las observaciones en un par tienen algo en común. Para determinar si la dieta es efectiva consideramos las diferencias $d_1, d_2, \ldots, d_n$ en las observaciones pareadas. Estas diferencias son los valores de una muestra aleatoria $D_1, D_2, \ldots, D_n$ de una población de diferencias, que supondremos distribuidas normalmente, con media $\mu_D = \mu_1 - \mu_2$ y varianza σ_D^2. Estimamos σ_D^2 mediante s_d^2, la varianza de las diferencias que constituyen nuestra muestra. El estimador puntual de μ_D es dado por $\bar{D}$.

¿Cuándo debe hacerse el pareado?

Parear observaciones en un experimento es una estrategia que se puede emplear en muchos campos de aplicación. Se expondrá al lector a tal concepto en el material relacionado con

la prueba de hipótesis en el capítulo 10 y en los temas de diseño experimental en los capítulos 13 y 15. Al seleccionar unidades experimentales relativamente homogéneas (dentro de las unidades) y permitir que cada unidad experimente ambas condiciones de la población, se reduce la varianza del error experimental efectiva (en este caso σ_D^2). El lector puede visualizar la *i*-ésima diferencia del par como

$$D_i = X_{1i} - X_{2i}.$$

Como las dos observaciones se toman de la unidad experimental de la muestra no son independientes y, de hecho,

$$\text{Var}(D_i) = \text{Var}(X_{1i} - X_{2i}) = \sigma_1^2 + \sigma_2^2 - 2\,\text{Cov}(X_{1i}, X_{2i}).$$

Entonces, de manera intuitiva, se espera que σ_D^2 debería reducirse debido a la similitud en la naturaleza de los "errores" de las dos observaciones dentro de una unidad experimental, a lo cual se llega mediante la expresión anterior. En realidad se espera que, si la unidad es homogénea, la covarianza sea positiva. Como resultado, la ganancia en calidad del intervalo de confianza sobre la que se obtuvo sin parear es mayor cuando hay homogeneidad dentro de las unidades y cuando las diferencias grandes van de una a otra unidad. Se debería tener en cuenta que el desempeño del intervalo de confianza dependerá del error estándar de $\bar{D}$, que es, por supuesto, $\sigma_D/\sqrt{n}$, donde n es el número de pares. Como indicamos antes, la intención al parear es reducir σ_D.

Equilibrio entre reducir la varianza y perder grados de libertad

Al comparar los intervalos de confianza obtenidos con y sin pareado es evidente que hay un intercambio implicado. Aunque en realidad el pareado debería reducir la varianza y, por lo tanto, el error estándar de la estimación puntual, los grados de libertad disminuyen al reducir el problema a uno con una sola muestra. Como resultado, el punto $t_{\alpha/2}$ ligado al error estándar se ajusta en concordancia. De esta manera, el pareado podría resultar contraproducente. Esto ocurriría con certeza si se experimenta sólo una reducción modesta en la varianza (a través de σ_D^2) mediante el pareado.

Otra ilustración del pareado implicaría elegir n pares de sujetos, donde cada par tenga una característica similar, como el coeficiente intelectual (CI), la edad o la raza, y luego para cada par seleccionar un miembro al azar para obtener un valor de X_1, dejando que el otro miembro proporcione el valor de X_2. En este caso, X_1 y X_2 podrían representar las calificaciones obtenidas por dos individuos con igual CI cuando uno es asignado al azar a un grupo que usa el método de enseñanza convencional y al otro a un grupo que utiliza materiales programados.

Se puede establecer un intervalo de confianza del $100(1-\alpha)\%$ para μ_D escribiendo

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}}$ y $t_{\alpha/2}$, como antes, es un valor de la distribución t con $n-1$ grados de libertad.

En la actualidad se acostumbra reemplazar T por su definición en la desigualdad anterior y desarrollar los pasos matemáticos que conduzcan al siguiente intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2 = \mu_D$.

Intervalo de confianza para $\mu_D = \mu_1 - \mu_2$ para observaciones pareadas

Si $\bar{d}$ y s_d son la media y la desviación estándar, respectivamente, de las diferencias distribuidas normalmente de n pares aleatorios de mediciones, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_D = \mu_1 - \mu_2$ es

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad, que deja una área de $\alpha/2$ a la derecha.

Ejemplo 9.13: Un estudio publicado en *Chemosphere* reporta los niveles de la dioxina TCDD en 20 veteranos de Vietnam de Massachusetts, quienes posiblemente estuvieron expuestos al agente naranja. En la tabla 9.1 se presentan los niveles de TCDD en plasma y tejido adiposo.

Calcule un intervalo de confianza del 95% para $\mu_1 - \mu_2$, donde μ_1 y μ_2 representen las medias verdaderas de los niveles de TCDD en plasma y en tejido adiposo, respectivamente. Suponga que la distribución de las diferencias es casi normal.

Tabla 9.1: Datos para el ejemplo 9.13.

Veterano	Niveles de TCDD en plasma	Niveles de TCDD en tejido adiposo	d_i	Veterano	Niveles de TCDD en plasma	Niveles de TCDD en tejido adiposo	d_i
1	2.5	4.9	−2.4	11	6.9	7.0	−0.1
2	3.1	5.9	−2.8	12	3.3	2.9	0.4
3	2.1	4.4	−2.3	13	4.6	4.6	0.0
4	3.5	6.9	−3.4	14	1.6	1.4	0.2
5	3.1	7.0	−3.9	15	7.2	7.7	−0.5
6	1.8	4.2	−2.4	16	1.8	1.1	0.7
7	6.0	10.0	−4.0	17	20.0	11.0	9.0
8	3.0	5.5	−2.5	18	2.0	2.5	−0.5
9	36.0	41.0	−5.0	19	2.5	2.3	0.2
10	4.7	4.4	0.3	20	4.1	2.5	1.6

Reproducido de *Chemosphere*, Vol. 20, Núms. 7-9 (tablas I y II), Schecter *et al*., "Partitioning 2, 3, 7, 8-chlorinated dibenzo-p-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans", pp. 954-955, Derechos reservados ©1990, con autorización de Elsevier.

***Solución*:** Buscamos un intervalo de confianza del 95% para $\mu_1 - \mu_2$. Como las observaciones están pareadas, $\mu_1 - \mu_2 = \mu_D$. La estimación puntual de μ_D es $\bar{d} = -0.87$. La desviación estándar s_d de las diferencias muestrales es

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{168.4220}{19}} = 2.9773.$$

Si usamos $\alpha = 0.05$, en la tabla A.4 encontramos que $t_{0.025} = 2.093$ para $v = n - 1 = 19$ grados de libertad. Por lo tanto, el intervalo de confianza del 95% es

$$-0.8700 - (2.093)\left(\frac{2.9773}{\sqrt{20}}\right) < \mu_D < -0.8700 + (2.093)\left(\frac{2.9773}{\sqrt{20}}\right),$$

o simplemente $-2.2634 < \mu_D < 0.5234$, de lo cual concluimos que no hay diferencia significativa entre el nivel medio de TCDD en plasma y el nivel medio de TCDD en tejido adiposo.

Ejercicios

9.35 Una muestra aleatoria de tamaño $n_1 = 25$, tomada de una población normal con una desviación estándar $\sigma_1 = 5$, tiene una media $\bar{x}_1 = 80$. Una segunda muestra aleatoria de tamaño $n_2 = 36$, que se toma de una población normal diferente con una desviación estándar $\sigma_2 = 3$, tiene una media $\bar{x}_2 = 75$. Calcule un intervalo de confianza del 94% para $\mu_1 - \mu_2$.

9.36 Se comparan las resistencias de dos clases de hilo. Se prueban 50 piezas de cada clase de hilo en condiciones similares. La marca *A* tiene una resistencia a la tensión promedio de 78.3 kilogramos, con una desviación estándar de 5.6 kilogramos; en tanto que la marca *B* tiene una resistencia a la tensión promedio de 87.2 kilogramos con una desviación estándar de 6.3 kilogramos. Construya un intervalo de confianza del 95% para la diferencia de las medias de la población.

9.37 Se realiza un estudio para determinar si cierto tratamiento tiene algún efecto sobre la cantidad de metal que se elimina en una operación de encurtido. Una muestra aleatoria de 100 piezas se sumerge en un baño por 24 horas sin el tratamiento, lo que produce un promedio de 12.2 milímetros de metal eliminados y una desviación estándar muestral de 1.1 milímetros. Una segunda muestra de 200 piezas se somete al tratamiento, seguido de 24 horas de inmersión en el baño, lo que da como resultado una eliminación promedio de 9.1 milímetros de metal, con una desviación estándar muestral de 0.9 milímetros. Calcule un estimado del intervalo de confianza del 98% para la diferencia entre las medias de las poblaciones. ¿El tratamiento parece reducir la cantidad media del metal eliminado?

9.38 En un proceso químico por lotes se comparan los efectos de dos catalizadores sobre la potencia de la reacción del proceso. Se prepara una muestra de 12 lotes utilizando el catalizador 1 y una muestra de 10 lotes utilizando el catalizador 2. Los 12 lotes para los que se utilizó el catalizador 1 en la reacción dieron un rendimiento promedio de 85 con una desviación estándar muestral de 4; en tanto que para la segunda muestra, la de 10 lotes, el promedio fue de 81, con una desviación estándar muestral de 5. Calcule un intervalo de confianza del 90% para la diferencia entre las medias de la población, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal y que tienen varianzas iguales.

9.39 Los estudiantes pueden elegir entre un curso de física de tres semestres-hora sin laboratorio y un curso de cuatro semestres-hora con laboratorio. El examen final escrito es el mismo para ambos cursos. Si 12 estudiantes del curso con laboratorio obtienen una calificación promedio de 84, con una desviación estándar de 4, y 18 estudiantes del grupo sin laboratorio obtienen una calificación promedio de 77, con una desviación estándar de 6, calcule un intervalo de confianza del 99% para la diferencia entre las calificaciones promedio para ambos cursos. Suponga que las poblaciones se distribuyen de forma aproximadamente normal y que tienen varianzas iguales.

9.40 En un estudio que se lleva a cabo en Virginia Tech sobre el desarrollo de micorriza, una relación simbiótica entre las raíces de árboles y un hongo, en la cual se transfieren minerales del hongo a los árboles y azúcares de los árboles a los hongos, se cultivaron en un invernadero 20 robles rojos que fueron expuestos al hongo *Pisolithus tinctorus*. Todos los árboles se plantaron en el mismo tipo de suelo y recibieron la misma cantidad de luz solar y agua. La mitad no recibió nitrógeno en el momento de plantarlos y sirvió como control, y la otra mitad recibió 368 ppm de nitrógeno en forma de $NaNO_3$. Después de 140 días se registraron los siguientes pesos de los tallos, en gramos:

Sin nitrógeno	Con nitrógeno
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

Construya un intervalo de confianza del 95% para la diferencia entre los pesos medios de los tallos que no recibieron nitrógeno y los que recibieron 368 ppm de nitrógeno. Suponga que las poblaciones están distribuidas normalmente y que tienen varianzas iguales.

9.41 Los siguientes datos representan el tiempo, en días, que pacientes tratados al azar con uno de dos medicamentos para curar infecciones graves de la vejiga tardaron en recuperarse:

Medicamento 1	Medicamento 2
$n_1 = 14$	$n_2 = 16$
$\bar{x}_1 = 17$	$\bar{x}_2 = 19$
$s_1^2 = 1.5$	$s_2^2 = 1.8$

Calcule un intervalo de confianza del 99% para la diferencia $\mu_2 - \mu_1$ en los tiempos medios de recuperación para los dos medicamentos. Suponga poblaciones normales que tienen varianzas iguales.

9.42 Un experimento publicado en *Popular Science* comparó el ahorro de combustible para dos tipos de camiones compactos que funcionan con diesel y están equipados de forma similar. Suponga que se utilizaron 12 camiones Volkswagen y 10 Toyota en pruebas con una velocidad constante de 90 kilómetros por hora. Si los 12 camiones Volkswagen promedian 16 kilómetros por litro con una desviación estándar de 1.0 kilómetros por litro, y los 10 Toyota promedian 11 kilómetros por litro con una desviación estándar de 0.8 kilómetros por litro, construya un intervalo de confianza del 90% para la diferencia entre los kilómetros promedio por litro de estos dos camiones compactos. Suponga que las distancias por litro para cada modelo de camión están distribuidas de forma aproximadamente normal y que tienen varianzas iguales.

9.43 Una empresa de taxis trata de decidir si comprará neumáticos de la marca *A* o de la marca *B* para su flotilla de taxis. Para estimar la diferencia entre las dos marcas realiza un experimento utilizando 12 neumáticos de cada marca, los cuales utiliza hasta que se desgastan. Los resultados son:

Marca *A*: $\bar{x}_1 = 36{,}300$ kilómetros, $s_1 = 5000$ kilómetros.
Marca *B*: $\bar{x}_2 = 38{,}100$ kilómetros, $s_2 = 6100$ kilómetros.

Calcule un intervalo de confianza del 95% para $\mu_A - \mu_B$, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal. Puede no suponer que las varianzas son iguales.

9.44 Con referencia al ejercicio 9.43, calcule un intervalo de confianza del 99% para $\mu_1 - \mu_2$ si se asignan al azar neumáticos de las dos marcas a las ruedas traseras izquierda y derecha de 8 taxis y se registran las siguientes distancias, en kilómetros:

Taxi	Marca *A*	Marca *B*
1	34,400	36,700
2	45,500	46,800
3	36,700	37,700
4	32,000	31,100
5	48,400	47,800
6	32,800	36,400
7	38,100	38,900
8	30,100	31,500

Suponga que las diferencias de las distancias se distribuyen de forma aproximadamente normal.

9.45 El gobierno otorgó fondos para los departamentos de agricultura de 9 universidades para probar las capacidades de cosecha de dos nuevas variedades de trigo. Cada variedad se siembra en parcelas con la misma área en cada universidad, y las cosechas, en kilogramos por parcela, se registran como sigue:

	Universidad								
Variedad	1	2	3	4	5	6	7	8	9
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

Calcule un intervalo de confianza del 95% para la diferencia media entre las cosechas de las dos variedades, suponiendo que las diferencias entre las cosechas se distribuyen de forma aproximadamente normal. Explique por qué es necesario el pareado en este problema.

9.46 Los siguientes datos representan el tiempo de duración de películas producidas por dos empresas cinematográficas.

Empresa	Tiempo (minutos)						
I	103	94	110	87	98		
II	97	82	123	92	175	88	118

Calcule un intervalo de confianza del 90% para la diferencia entre la duración promedio de las películas que producen las dos empresas. Suponga que las diferencias en la duración se distribuyen de forma aproximadamente normal y que tienen varianzas distintas.

9.47 La revista *Fortune* (marzo de 1997) publicó la rentabilidad total de los inversionistas durante los 10 años anteriores a 1996 y también la de 431 empresas en ese mismo año. A continuación se lista la rentabilidad total para 10 de las empresas. Calcule un intervalo de confianza del 95% para el cambio promedio en el porcentaje de rentabilidad de los inversionistas.

	Rentabilidad total para los inversionistas	
Empresa	**1986–96**	**1996**
Coca-Cola	29.8%	43.3%
Mirage Resorts	27.9%	25.4%
Merck	22.1%	24.0%
Microsoft	44.5%	88.3%
Johnson & Johnson	22.2%	18.1%
Intel	43.8%	131.2%
Pfizer	21.7%	34.0%
Procter & Gamble	21.9%	32.1%
Berkshire Hathaway	28.3%	6.2%
S&P 500	11.8%	20.3%

9.48 Una empresa automotriz está considerando dos tipos de baterías para sus vehículos. Con ese fin reúne información muestral sobre la vida de las baterías. Utiliza para ello 20 baterías del tipo *A* y 20 baterías del tipo *B*. El resumen de los estadísticos es $\bar{x}_A = 32.91$,

$\bar{x}_B = 30.47$, $s_A = 1.57$ y $s_B = 1.74$. Suponga que los datos de cada batería se distribuyen normalmente y que $\sigma_A = \sigma_B$.

a) Calcule un intervalo de confianza del 95% para $\mu_A - \mu_B$.

b) Del inciso *a*) saque algunas conclusiones que le ayuden a la empresa a decidir si debería utilizar la batería *A* o la *B*.

9.49 Se considera usar dos marcas diferentes de pintura vinílica. Se seleccionaron 15 especímenes de cada tipo de pintura, para los cuales los tiempos de secado en horas fueron los siguientes:

Pintura *A*	**Pintura *B***
3.5 2.7 3.9 4.2 3.6	4.7 3.9 4.5 5.5 4.0
2.7 3.3 5.2 4.2 2.9	5.3 4.3 6.0 5.2 3.7
4.4 5.2 4.0 4.1 3.4	5.5 6.2 5.1 5.4 4.8

Suponga que el tiempo de secado se distribuye normalmente, con $\sigma_A = \sigma_B$. Calcule un intervalo de confianza del 95% de $\mu_B - \mu_A$, donde μ_A y μ_B son los tiempos medios de secado.

9.50 A dos grupos de ratas diabéticas se les suministran dos niveles de dosis de insulina (alto y bajo) para verificar la capacidad de fijación de esta hormona. Se obtuvieron los siguientes datos.

Dosis baja: $n_1 = 8$ $\bar{x}_1 = 1.98$ $s_1 = 0.51$
Dosis alta: $n_2 = 13$ $\bar{x}_2 = 1.30$ $s_2 = 0.35$

Suponga que las varianzas son iguales. Determine un intervalo de confianza del 95% para la diferencia en la capacidad promedio verdadera de fijación de la insulina entre las dos muestras.

9.10 Una sola muestra: estimación de una proporción

El estadístico $\hat{P} = X/n$, en donde X representa el número de éxitos en n ensayos, provee un estimador puntual de la proporción p en un experimento binomial. Por lo tanto, la proporción de la muestra $\hat{p} = x/n$ se utilizará como el estimador puntual del parámetro p.

Si no se espera que la proporción p desconocida esté demasiado cerca de 0 o de 1, se puede establecer un intervalo de confianza para p considerando la distribución muestral de $\hat{P}$. Si en cada ensayo binomial asignamos el valor 0 a un fracaso y el valor 1 a un éxito, el número de éxitos, x, se puede interpretar como la suma de n valores que consta sólo de ceros y unos, y $\hat{p}$ es sólo la media muestral de esos n valores. En consecuencia, por el teorema del límite central, para n suficientemente grande $\hat{P}$ está distribuida de forma casi normal con media

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

y varianza

$$\sigma^2_{\hat{P}} = \sigma^2_{X/n} = \frac{\sigma^2_X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}.$$

Por lo tanto, podemos afirmar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \text{ con } Z = \frac{\hat{P} - p}{\sqrt{pq/n}},$$

y $z_{\alpha/2}$ es el valor por arriba del cual encontramos una área de $\alpha/2$ debajo de la curva normal estándar. Al sustituir para Z escribimos

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Cuando n es grande se introduce un error muy pequeño sustituyendo el estimado puntual $\hat{p} = x/n$ para la p debajo del signo de radical. Entonces podemos escribir

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha.$$

Por otro lado, al resolver para p en la desigualdad cuadrática anterior,

$$-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2},$$

obtenemos otra forma del intervalo de confianza para p con los siguientes límites:

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

Para una muestra aleatoria de tamaño n se calcula la proporción muestral $\hat{p} = x/n$ y se pueden obtener los siguientes intervalos de confianza aproximados del $100(1-\alpha)\%$ para p.

Intervalos de confianza para p de una muestra grande

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para el parámetro binomial p se obtiene por medio de (método 1)

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

o mediante (método 2)

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} - \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} < p < \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} + \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Cuando n es pequeña y se cree que la proporción desconocida p se acerca a 0 o a 1, el procedimiento del intervalo de confianza que se establece aquí no es confiable y, por lo tanto, no se debería emplear. Para estar seguros se requiere que tanto $n\hat{p}$ como $n\hat{q}$ sean mayores que o iguales a 5. Los métodos para calcular un intervalo de confianza para el parámetro binomial p también se pueden aplicar cuando se está utilizando la distribución binomial con el fin de aproximar la distribución hipergeométrica; es decir, cuando n es pequeña respecto a N, como se ilustra en el ejemplo 9.14.

Observe que, aunque el método 2 produce resultados más precisos, su cálculo es más complicado, y la ventaja en precisión que brinda disminuye cuando el tamaño de la muestra es lo suficientemente grande. Debido a esto en la práctica es más común utilizar el método 1.

Ejemplo 9.14: En una muestra aleatoria de $n = 500$ familias que tienen televisores en la ciudad de Hamilton, Canadá, se encuentra que $x = 340$ están suscritas a HBO. Calcule un intervalo de confianza del 95% para la proporción real de familias que tienen televisores en esta ciudad y están suscritas a HBO.

***Solución*:** La estimación puntual de p es $\hat{p} = 340/500 = 0.68$. Si usamos la tabla A.3, encontramos que $z_{0.025} = 1.96$. Por lo tanto, si utilizamos el método 1, el intervalo de confianza del 95% para p es

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}},$$

que se simplifica a $0.6391 < p < 0.7209$.

Si utilizamos el segundo método, obtenemos

$$\frac{0.68 + \frac{1.96^2}{(2)(500)}}{1 + \frac{1.96^2}{500}} \pm \frac{1.96}{1 + \frac{1.96^2}{500}} \sqrt{\frac{(0.68)(0.32)}{500} + \frac{1.96^2}{(4)(500^2)}} = 0.6786 \pm 0.0408,$$

que se simplifica a $0.6378 < p < 0.7194$. Aparentemente, cuando n es grande (500 en este caso) ambos métodos producen resultados muy similares.

Si p es el valor central de un intervalo de confianza del $100(1-\alpha)\%$, entonces $\hat{p}$ estima p sin error. Sin embargo, la mayoría de las veces $\hat{p}$ no será exactamente igual a p y el estimado puntual será erróneo. El tamaño de este error será la diferencia positiva que separa a p de $\hat{p}$, y podemos tener una confianza del $100(1-\alpha)\%$ de que tal diferencia no excederá a $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$. Si dibujamos un diagrama de un intervalo de confianza típico, como el de la figura 9.6, podemos ver esto fácilmente. En este caso utilizamos el método 1 para estimar el error.

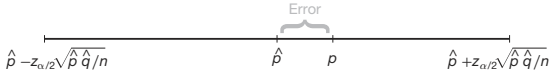


Figura 9.6: Error en la estimación de p por medio de $\hat{p}$.

Teorema 9.3: Si $\hat{p}$ se utiliza como un estimado de p, podemos tener un $100(1-\alpha)\%$ de confianza en que el error no excederá a $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

En el ejemplo 9.14 tenemos un 95% de confianza en que la proporción de la muestra $\hat{p} = 0.68$ difiere de la verdadera proporción p en una cantidad que no excede a 0.04.

Selección del tamaño de la muestra

Determinemos ahora qué tan grande debe ser una muestra para poder estar seguros de que el error al estimar p será menor que una cantidad específica e. Por medio del teorema 9.3, debemos elegir una n tal que $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} = e$.

Teorema 9.4: Si $\hat{p}$ se utiliza como un estimado de p, podemos tener un $100(1-\alpha)\%$ de confianza en que el error será menor que una cantidad específica e cuando el tamaño de la muestra sea aproximadamente

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}.$$

El teorema 9.4 es algo engañoso, pues debemos utilizar $\hat{p}$ para determinar el tamaño n de la muestra, pero $\hat{p}$ se calcula a partir de la muestra. Si se puede hacer una estimación burda de p sin tomar una muestra, se podría usar este valor para determinar n. A falta de tal estimado, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ para proporcionar un estimado de p. Si utilizamos el teorema 9.4 podríamos determinar aproximadamente cuántas observaciones se necesitan para proporcionar el grado de precisión deseado. Observe que los valores fraccionarios de n se redondean al siguiente número entero mayor.

Ejemplo 9.15: ¿Qué tan grande debe ser una muestra en el ejemplo 9.14 si queremos tener un 95% de confianza en que la estimación de p esté dentro de 0.02 del valor verdadero?

***Solución*:** Tratemos a las 500 familias como una muestra preliminar que proporciona una estimación $\hat{p} = 0.68$. Entonces, mediante el teorema 9.4,

$$n = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.8 \approx 2090.$$

Por lo tanto, si basamos nuestra estimación de p en una muestra aleatoria de tamaño 2090, podemos tener un 95% de confianza en que nuestra proporción muestral no diferirá de la proporción verdadera en más de 0.02. ∎

Ocasionalmente será poco práctico obtener una estimación de p que se utilice para determinar el tamaño muestral para un grado específico de confianza. Si esto sucede, se establece un límite superior para n al notar que $\hat{p}\hat{q} = \hat{p}(1 - \hat{p})$, que debe ser a lo sumo $1/4$, ya que $\hat{p}$ debe caer entre 0 y 1. Este hecho se verifica completando el cuadrado. Por consiguiente,

$$\hat{p}(1 - \hat{p}) = -(\hat{p}^2 - \hat{p}) = \frac{1}{4} - \left(\hat{p}^2 - \hat{p} + \frac{1}{4}\right) = \frac{1}{4} - \left(\hat{p} - \frac{1}{2}\right)^2,$$

que siempre es menor que $1/4$ excepto cuando $\hat{p} = 1/2$ y entonces $\hat{p}\hat{q} = 1/4$. Por lo tanto, si sustituimos $\hat{p} = 1/2$ en la fórmula para n del teorema 9.4, cuando, de hecho, p difiere de $1/2$, entonces n se agrandará más de lo necesario para el grado de confianza específico y, como resultado, se incrementará nuestro grado de confianza.

Teorema 9.5: Si utilizamos $\hat{p}$ como un estimado de p, podemos tener, **al menos**, un $100(1 - \alpha)\%$ de confianza en que el error no excederá a una cantidad específica e cuando el tamaño de la muestra sea

$$n = \frac{z_{\alpha/2}^2}{4e^2}.$$

Ejemplo 9.16: ¿Qué tan grande debe ser una muestra en el ejemplo 9.14 si queremos tener al menos un 95% de confianza en que nuestra estimación de p está dentro de 0.02 del valor verdadero?

***Solución*:** A diferencia del ejemplo 9.15, supondremos ahora que no se tomó una muestra preliminar para obtener una estimación de p. En consecuencia, podemos tener al menos un 95% de confianza en que nuestra proporción de la muestra no diferirá de la proporción verdadera en más de 0.02, si elegimos una muestra de tamaño

$$n = \frac{(1.96)^2}{(4)(0.02)^2} = 2401.$$

Si comparamos los resultados de los ejemplos 9.15 y 9.16, vemos que la información concerniente a p, proporcionada por una muestra preliminar, o quizás obtenida a partir de la experiencia, nos permite elegir una muestra más pequeña a la vez que mantenemos el grado de precisión requerido. ∎

9.11 Dos muestras: estimación de la diferencia entre dos proporciones

Considere el problema en el que se busca estimar la diferencia entre dos parámetros binomiales p_1 y p_2. Por ejemplo, p_1 podría ser la proporción de fumadores con cáncer de pulmón y p_2 la proporción de no fumadores con cáncer de pulmón, y el problema consistiría en estimar la diferencia entre estas dos proporciones. Primero seleccionamos muestras aleatorias independientes de tamaños n_1 y n_2 a partir de las dos poblaciones binomiales con medias n_1p_1 y n_2p_2, y varianzas $n_1p_1q_1$ y $n_2p_2q_2$, respectivamente, después determinamos los números x_1 y x_2 de personas con cáncer de pulmón en cada muestra, y formamos las proporciones $\hat{p}_1 = x_1/n$ y $\hat{p}_2 = x_2/n$. El estadístico $\hat{P}_1 - \hat{P}_2$ provee un estimador puntual de la diferencia entre las dos proporciones, $p_1 - p_2$. Por lo tanto, la diferencia de las proporciones muestrales, $\hat{p}_1 - \hat{p}_2$, se utilizará como la estimación puntual de $p_1 - p_2$.

Se puede establecer un intervalo de confianza para $p_1 - p_2$ considerando la distribución muestral de $\hat{P}_1 - \hat{P}_2$. De la sección 9.10 sabemos que $\hat{P}_1$ y $\hat{P}_2$ están distribuidos cada uno de forma aproximadamente normal, con medias p_1 y p_2, y varianzas p_1q_1/n_1 y p_2q_2/n_2, respectivamente. Al elegir muestras independientes de las dos poblaciones nos aseguramos de que las variables $\hat{P}_1$ y $\hat{P}_2$ serán independientes y luego, por la propiedad reproductiva de la distribución normal que se estableció en el teorema 7.11, concluimos que $\hat{P}_1 - \hat{P}_2$ está distribuido de forma aproximadamente normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}.$$

Por lo tanto, podemos asegurar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}}$$

y $z_{\alpha/2}$ es un valor por arriba del cual encontramos una área de $\alpha/2$ debajo de la curva normal estándar. Al sustituir para Z escribimos

$$P\left[-z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Después de realizar las operaciones matemáticas usuales reemplazamos p_1, p_2, q_1 y q_2 bajo el signo de radical por sus estimaciones $\hat{p}_1 = x_1/n_1$, $\hat{p}_2 = x_2/n_2$, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, siempre y cuando $n_1\hat{p}_1$, $n_2\hat{q}_1$, $n_2\hat{p}_2$ y $n_1\hat{q}_1$ sean todas mayores que o iguales a 5, y se obtiene el siguiente intervalo de confianza aproximado del $100(1-\alpha)\%$ para $p_1 - p_2$.

Intervalo de confianza para $p_1 - p_2$ de una muestra grande

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatorias de tamaños n_1 y n_2, respectivamente, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$ es dado por

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Ejemplo 9.17: Se considera hacer un cierto cambio en el proceso de fabricación de partes componentes. Para determinar si el cambio en el proceso da como resultado una mejora, se toman muestras de partes fabricadas con el proceso nuevo y con el actual. Si se encuentra que 75 de 1500 artículos manufacturados con el proceso actual están defectuosos y 80 de 2000 manufacturados con el proceso nuevo también lo están, calcule un intervalo de confianza del 90% para la diferencia verdadera en la proporción de partes defectuosas entre el proceso actual y el nuevo.

***Solución*:** Suponga que p_1 y p_2 son las proporciones verdaderas de partes defectuosas para los procesos actual y nuevo, respectivamente. En consecuencia, $\hat{p}_1 = 75/1500 = 0.05$ y $\hat{p}_2 = 80/2000 = 0.04$, y la estimación puntual de $p_1 - p_2$ es

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01.$$

Si utilizamos la tabla A.3, encontramos $z_{0.05} = 1.645$. Por lo tanto, al sustituir en la fórmula

$$1.645\sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}} = 0.0117,$$

encontramos que el intervalo de confianza del 90% es $-0.0017 < p_1 - p_2 < 0.0217$. Como el intervalo contiene el valor 0, no hay razón para creer que el nuevo proceso, comparado con el actual, disminuye en forma significativa la proporción de artículos defectuosos. ■

Hasta aquí todos los intervalos de confianza presentados son de la forma

$$\text{estimación puntual} \pm K \text{ e.e.(estimación puntual)},$$

donde K es una constante (ya sea t o el punto porcentual normal). Esta forma es válida cuando el parámetro es una media, una diferencia entre medias, una proporción o una diferencia entre proporciones, debido a la simetría de las distribuciones t y Z. Sin embargo, no se extiende a las varianzas ni a los cocientes de las varianzas, las cuales se examinarán en las secciones 9.12 y 9.13.

Ejercicios

En este conjunto de ejercicios, para una estimación respecto a una proporción, utilice sólo el método 1 para calcular los intervalos de confianza, a menos que se especifique otra cosa.

9.51 En una muestra aleatoria de 1000 viviendas en cierta ciudad se encuentra que 228 utilizan petróleo como combustible para la calefacción. Calcule intervalos de confianza del 99% para la proporción de viviendas en esta ciudad que utilizan petróleo con el fin mencionado. Utilice los dos métodos que se presentaron en la página 297.

9.52 Calcule intervalos de confianza del 95% para la proporción de artículos defectuosos que resultan de un proceso cuando se encuentra que una muestra de tamaño 100 produce 8 defectuosos. Utilice los dos métodos que se presentaron en la página 297.

9.53 *a*) Se selecciona una muestra aleatoria de 200 votantes en una ciudad y se encuentra que 114 apoyan un juicio de anexión. Calcule el intervalo de confianza del 96% para la parte de la población votante que está a favor del juicio.

b) ¿Qué podemos afirmar con 96% de confianza acerca de la posible magnitud de nuestro error, si estimamos que la fracción de votantes que está a favor del juicio de anexión es 0.57?

9.54 Un fabricante de reproductores de MP3 utiliza un conjunto de pruebas exhaustivas para evaluar el funcionamiento eléctrico de su producto. Todos los reproductores de MP3 deben pasar todas las pruebas antes de ser puestos a la venta. De una muestra aleatoria de 500 reproductores, 15 no pasan una o más de las pruebas. Calcule un intervalo de confianza del 90% para la proporción de los reproductores de MP3 de la población que pasan todas las pruebas.

9.55 Se está considerando un nuevo sistema de lanzamiento de cohetes para el despliegue de cohetes pequeños, de corto alcance. La probabilidad de que el sistema existente tenga un lanzamiento exitoso se representa con $p = 0.8$. Se toma una muestra de 40 lanzamientos experimentales con el nuevo sistema y 34 resultan exitosos.

a) Construya un intervalo de confianza del 95% para p.

b) ¿Con base en sus resultados, concluiría que el nuevo sistema es mejor?

9.56 Un genetista está interesado en determinar la proporción de hombres africanos que padecen cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres africanos encuentra que 24 lo padecen.

a) Calcule un intervalo de confianza del 99% para la proporción de hombres africanos que padecen este trastorno sanguíneo.

b) ¿Qué podríamos afirmar con 99% de confianza acerca de la posible magnitud de nuestro error, si estimamos que la proporción de hombres africanos con dicho trastorno sanguíneo es 0.24?

9.57 *a*) De acuerdo con un reporte del *Roanoke Times & World-News*, aproximadamente $2/3$ de los 1600 adultos encuestados vía telefónica dijeron que piensan que invertir en el programa del transbordador espacial es bueno para Estados Unidos. Calcule un intervalo de confianza del 95% para la proporción de adultos estadounidenses que piensan que el programa del transbordador espacial es una buena inversión para su país.

b) ¿Qué podríamos afirmar con un 95% de confianza acerca de la posible magnitud de nuestro error, si estimamos que la proporción de adultos estadounidenses que piensan que el programa del transbordador espacial es una buena inversión es de $2/3$?

9.58 En el artículo del periódico al que se hace referencia en el ejercicio 9.57, 32% de los 1600 adultos encuestados dijo que el programa espacial estadounidense debería enfatizar la exploración científica. ¿Qué tamaño debería tener una muestra de adultos para la encuesta si se desea tener un 95% de confianza en que el porcentaje estimado esté dentro del 2% del porcentaje verdadero?

9.59 ¿Qué tamaño debería tener una muestra si deseamos tener un 96% de confianza en que nuestra proporción de la muestra en el ejercicio 9.53 esté dentro del 0.02 de la fracción verdadera de la población votante?

9.60 ¿Qué tamaño debería tener una muestra si deseamos tener un 99% de confianza en que nuestra proporción de la muestra en el ejercicio 9.51 esté dentro del 0.05 de la proporción verdadera de viviendas en esa ciudad que utilizan petróleo como combustible para la calefacción?

9.61 ¿Qué tamaño debería tener una muestra en el ejercicio 9.52 si deseamos tener un 98% de confianza en que nuestra proporción de la muestra esté dentro del 0.05 de la proporción verdadera de defectuosos?

9.62 Una conjetura de un catedrático del departamento de microbiología, de la Facultad de Odontología de la Universidad de Washington, en St. Louis, Missouri, afirma que un par de tasas diarias de té verde o negro proporciona suficiente flúor para evitar el deterioro de los dientes. ¿Qué tan grande debería ser la muestra para estimar el porcentaje de habitantes de cierta ciudad que están a favor de tener agua fluorada, si se desea tener al menos un 99% de confianza en que el estimado está dentro del 1% del porcentaje verdadero?

9.63 Se llevará a cabo un estudio para estimar el porcentaje de ciudadanos de una ciudad que están a favor de tener agua fluorada. ¿Qué tan grande debería ser la muestra si se desea tener al menos 95% de confianza en que el estimado esté dentro del 1% del porcentaje verdadero?

9.64 Se realizará un estudio para estimar la proporción de residentes de cierta ciudad y sus suburbios que está a favor de que se construya una planta de energía nuclear cerca de la ciudad. ¿Qué tan grande debería ser la muestra, si se desea tener al menos un 95% de confianza en que el estimado esté dentro del 0.04 de la verdadera proporción de residentes que están a favor de que se construya la planta de energía nuclear?

9.65 A cierto genetista le interesa determinar la proporción de hombres y mujeres de la población que padecen cierto trastorno sanguíneo menor. En una muestra aleatoria de 1000 hombres encuentra que 250 lo padecen; mientras que de 1000 mujeres examinadas, 275 parecen padecerlo. Calcule un intervalo de confianza del 95% para la diferencia entre la proporción de hombres y mujeres que padecen el trastorno sanguíneo.

9.66 Se encuestan 10 escuelas de ingeniería de Estados Unidos. La muestra contiene a 250 ingenieros eléctricos, de los cuales 80 son mujeres; y 175 ingenieros químicos, de los cuales 40 son mujeres. Calcule un intervalo de confianza del 90% para la diferencia entre la proporción de mujeres en estos dos campos de la ingeniería. ¿Hay una diferencia significativa entre las dos proporciones?

9.67 Se llevó a cabo una prueba clínica para determinar si cierto tipo de vacuna tiene un efecto sobre la incidencia de cierta enfermedad. Una muestra de 1000 ratas, 500 de las cuales recibieron la vacuna, se mantuvo en un ambiente controlado durante un periodo de un año. En el grupo que no fue vacunado, 120 ratas presentaron la enfermedad, mientras que en el grupo inoculado 98 ratas la contrajeron. Si p_1 es la probabilidad de incidencia de la enfermedad en las ratas sin vacuna y p_2 es la probabilidad de incidencia en las ratas inoculadas, calcule un intervalo de confianza del 90% para $p_1 - p_2$.

9.68 En el estudio *Germination and Emergence of Broccoli,* realizado por el Departamento de horticultura del Virginia Tech, un investigador encontró que a 5°C, de 20 semillas de brócoli germinaron 10; en tanto que a 15°C, de 20 semillas germinaron 15. Calcule un intervalo de confianza del 95% para la diferencia en la proporción de semillas que germinaron a las dos temperaturas y decida si esta diferencia es significativa.

9.69 Una encuesta de 1000 estudiantes reveló que 274 eligen al equipo profesional de beisbol *A* como su equipo favorito. En 1991 se realizó una encuesta similar con 760 estudiantes y 240 de ellos también eligieron a ese equipo como su favorito. Calcule un intervalo de confianza del 95% para la diferencia entre la proporción de estudiantes que favorecen al equipo *A* en las dos encuestas. ¿Hay una diferencia significativa?

9.70 De acuerdo con el *USA Today* (17 de marzo de 1997), las mujeres constituían el 33.7% del personal de redacción en las estaciones locales de televisión en 1990 y el 36.2% en 1994. Suponga que en 1990 y en 1994 se contrataron 20 nuevos empleados para el personal de redacción.

a) Estime el número de trabajadores que habrían sido mujeres en 1990 y en 1994, respectivamente.

b) Calcule un intervalo de confianza del 95% para saber si hay evidencia de que la proporción de mujeres contratadas para el equipo de redacción fue mayor en 1994 que en 1990.

9.12 Una sola muestra: estimación de la varianza

Si extraemos una muestra de tamaño n de una población normal con varianza σ^2 y calculamos la varianza muestral s^2, obtenemos un valor del estadístico S^2. Esta varianza muestral calculada se utiliza como una estimación puntual de σ^2. En consecuencia, al estadístico S^2 se le denomina estimador de σ^2.

Se puede establecer una estimación por intervalos de σ^2 utilizando el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

De acuerdo con el teorema 8.4, cuando las muestras se toman de una población normal el estadístico X^2 tiene una distribución chi cuadrada con $n - 1$ grados de libertad. Podemos escribir (véase la figura 9.7)

$$P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha,$$

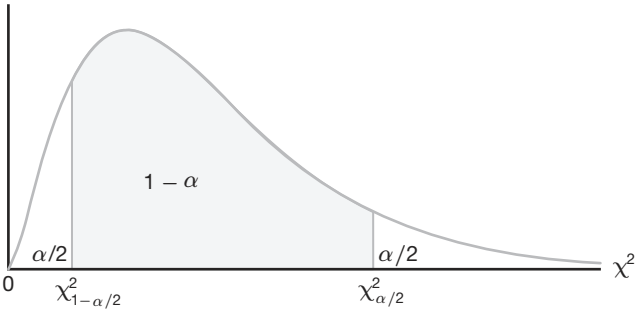


Figura 9.7: $P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$.

donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son valores de la distribución chi cuadrada con $n-1$ grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha. Al sustituir para X^2 escribimos

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1-\alpha.$$

Si dividimos cada término de la desigualdad entre $(n-1)S^2$, y después invertimos cada término (lo que cambia el sentido de las desigualdades), obtenemos

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha.$$

Para una muestra aleatoria de tamaño *n,* tomada de una población normal, se calcula la varianza muestral s^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ^2.

Intervalo de confianza para σ^2

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza del $100(1-\alpha)\%$ para σ^2 es

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son valores χ^2 con $v = n-1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente, a la derecha.

Un intervalo de confianza aproximado a $100(1-\alpha)\%$ para σ se obtiene tomando la raíz cuadrada de cada extremo del intervalo para σ^2.

Ejemplo 9.18: Los siguientes son los pesos, en decagramos, de 10 paquetes de semillas de pasto distribuidas por cierta empresa: 46.4, 46.1, 45.8, 47.0, 46.1. 45.9, 45.8, 46.9, 45.2 y 46.0. Calcule un intervalo de confianza del 95% para la varianza de todos los pesos de este tipo de paquetes de semillas de pasto distribuidos por la empresa. Suponga una población normal.

***Solución*:** Primero calculamos

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$

$$= \frac{(10)(21{,}273.12) - (461.2)^2}{(10)(9)} = 0.286.$$

Para obtener un intervalo de confianza del 95% elegimos $\alpha = 0.05$. Después, usando la tabla A.5 con $v = 9$ grados de libertad, encontramos $\chi^2_{0.025} = 19.023$ y $\chi^2_{0.975} = 2.700$. Por lo tanto, el intervalo de confianza del 95% para σ^2 es

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700},$$

o simplemente $0.135 < \sigma^2 < 0.953$.

9.13 Dos muestras: estimación de la proporción de dos varianzas

Una estimación puntual de la proporción de dos varianzas de la población σ_1^2/σ_2^2 es dada por la proporción s_1^2/s_2^2 de las varianzas muestrales. En consecuencia, el estadístico S_1^2/S_2^2 se conoce como un estimador de σ_1^2/σ_2^2.

Si σ_1^2 y σ_2^2 son las varianzas de poblaciones normales, podemos establecer una estimación por intervalos de σ_1^2/σ_2^2 usando el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

De acuerdo con el teorema 8.8, la variable aleatoria F tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, podemos escribir (véase la figura 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

donde $f_{1-\alpha/2}(v_1, v_2)$ y $f_{\alpha/2}(v_1, v_2)$ son los valores de la distribución F con v_1 y v_2 grados de libertad, que dejan áreas de $1 - \alpha/2$ y $\alpha/2$, respectivamente, a la derecha.

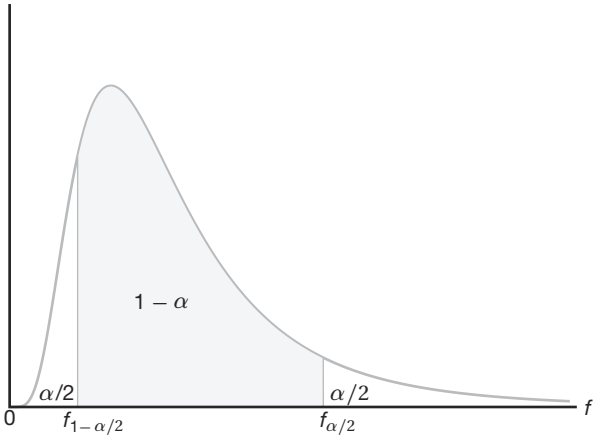


Figura 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

Al sustituir para *F,* escribimos

$$P\left[f_{1-\alpha/2}(v_1, v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha.$$

Si multiplicamos cada término de la desigualdad por S_2^2/S_1^2, y después invertimos cada término, obtenemos

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(v_1, v_2)}\right] = 1 - \alpha.$$

Los resultados del teorema 8.7 nos permiten reemplazar la cantidad $f_{1-\alpha/2}(v_1, v_2)$ por $1/f_{\alpha/2}(v_2, v_1)$. Por lo tanto,

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_2, v_1)\right] = 1 - \alpha.$$

Para cualesquiera dos muestras aleatorias independientes de tamaño n_1 y n_2 que se seleccionan de dos poblaciones normales, se calcula la proporción de las varianzas muestrales s_1^2/s_2^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2.

Intervalo de confianza para σ_1^2/σ_2^2

Si s_1^2 y s_2^2 son las varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, tomadas de poblaciones normales, entonces un intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2 es

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1),$$

donde $f_{\alpha/2}(v_1, v_2)$ es un valor f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha, y $f_{\alpha/2}(v_2, v_1)$ es un valor f similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad.

Como vimos en la sección 9.12, tomando la raíz cuadrada de cada extremo del intervalo para σ_1^2/σ_2^2, se obtiene un intervalo de confianza del $100(1-\alpha)\%$ para σ_1/σ_2.

Ejemplo 9.19: En el ejemplo 9.12 de la página 290 se construyó un intervalo de confianza para la diferencia en el contenido medio de ortofósforo de dos estaciones ubicadas sobre el río James, medido en miligramos por litro, suponiendo que las varianzas normales de la población son diferentes. Justifique esta suposición construyendo intervalos de confianza del 98% para σ_1^2/σ_2^2 y para σ_1/σ_2, donde σ_1^2 y σ_2^2 son las varianzas de la población del contenido de ortofósforo en la estación 1 y en la estación 2, respectivamente.

***Solución*:** Del ejemplo 9.12 tenemos $n_1 = 15$, $n_2 = 12$, $s_1 = 3.07$ y $s_2 = 0.80$. Para un intervalo de confianza del 98%, $\alpha = 0.02$. Al interpolar en la tabla A.6 encontramos $f_{0.01}(14,11) \approx 4.30$ y $f_{0.01}(11,14) \approx 3.87$. Por lo tanto, el intervalo de confianza del 98% para σ_1^2/σ_2^2 es

$$\left(\frac{3.07^2}{0.80^2}\right)\left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{3.07^2}{0.80^2}\right)(3.87),$$

que se simplifica a $3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$. Al calcular las raíces cuadradas de los límites de confianza encontramos que un intervalo de confianza del 98% para σ_1/σ_2 es

$$1.851 < \frac{\sigma_1}{\sigma_2} < 7.549.$$

Como este intervalo no permite la posibilidad de que σ_1/σ_2 sea igual a 1, es correcto suponer que $\sigma_1 \neq \sigma_2$ o $\sigma_1^2 \neq \sigma_2^2$ en el ejemplo 9.12. ■

Ejercicios

9.71 Un fabricante de baterías para automóvil afirma que sus baterías durarán, en promedio, 3 años con una varianza de 1 año. Suponga que 5 de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años y con base en esto construya un intervalo de confianza del 95% para σ^2, después decida si la afirmación del fabricante de que $\sigma^2 = 1$ es válida. Suponga que la población de duraciones de las baterías se distribuye de forma aproximadamente normal.

9.72 Una muestra aleatoria de 20 estudiantes obtuvo una media de $\bar{x} = 72$ y una varianza de $s^2 = 16$ en un examen universitario de colocación en matemáticas. Suponga que las calificaciones se distribuyen normalmente y con base en esto construya un intervalo de confianza del 98% para σ^2.

9.73 Construya un intervalo de confianza del 95% para σ^2 en el ejercicio 9.9 de la página 283.

9.74 Construya un intervalo de confianza del 99% para σ^2 en el ejercicio 9.11 de la página 283.

9.75 Construya un intervalo de confianza del 99% para σ en el ejercicio 9.12 de la página 283.

9.76 Construya un intervalo de confianza del 90% para σ en el ejercicio 9.13 de la página 283.

9.77 Construya un intervalo de confianza del 98% para σ_1/σ_2 en el ejercicio 9.42 de la página 295, donde σ_1 y σ_2 son, respectivamente, las desviaciones estándar para las distancias recorridas por litro de combustible de los camiones compactos Volkswagen y Toyota.

9.78 Construya un intervalo de confianza del 90% para σ_1^2/σ_2^2 en el ejercicio 9.43 de la página 295. ¿Se justifica que supongamos que $\sigma_1^2 \neq \sigma_2^2$ cuando construimos nuestro intervalo de confianza para $\mu_1 - \mu_2$?

9.79 Construya un intervalo de confianza del 90% para σ_1^2/σ_2^2 en el ejercicio 9.46 de la página 295. ¿Deberíamos suponer que $\sigma_1^2 = \sigma_2^2$ cuando construimos nuestro intervalo de confianza para $\mu_I - \mu_{II}$?

9.80 Construya un intervalo de confianza del 95% para σ_A^2/σ_B^2 en el ejercicio 9.49 de la página 295. ¿Tendría que utilizar la suposición de la igualdad de la varianza?

9.14 Estimación de la máxima verosimilitud (opcional)

A menudo los estimadores de parámetros han tenido que recurrir a la intuición. El estimador $\bar{X}$ ciertamente parece razonable como estimador de una media de la población μ. La virtud de S^2 como estimador de σ^2 se destaca en el estudio de estimadores insesgados de la sección 9.3. El estimador para un parámetro binomial p es simplemente una proporción de la muestra que, desde luego, es un *promedio* y recurre al sentido común. Sin embargo, hay muchas situaciones en las que no es del todo evidente cuál debería ser el estimador adecuado. Como resultado, el estudiante de estadística tiene mucho que aprender respecto a las diferentes filosofías que producen distintos métodos de estimación. En esta sección estudiaremos el **método de máxima verosimilitud**.

La estimación por máxima verosimilitud representa uno de los métodos de estimación más importantes en toda la estadística inferencial. No explicaremos el método de manera detallada; más bien, intentaremos transmitir la filosofía de la máxima verosimilitud e ilustrarla con ejemplos que la relacionan con otros problemas de estimación que se examinan en este capítulo.

Función de verosimilitud

Como el nombre lo indica, el método de máxima verosimilitud es aquel para el que se maximiza la *función de verosimilitud*, lo cual se ilustra mejor con un ejemplo que incluye una distribución discreta y un solo parámetro. Consideremos que $X_1, X_2, \ldots, X_n$ son las variables aleatorias independientes tomadas de una distribución de probabilidad discreta representada por $f(\mathbf{x}, \theta)$, donde θ es un solo parámetro de la distribución. Ahora bien,

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta) &= f(x_1, x_2, \ldots, x_n; \theta) \\ &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \end{aligned}$$

es la *distribución conjunta de las variables aleatorias,* la cual a menudo se denomina función de probabilidad. Observe que la variable de la función de probabilidad es θ, no $\mathbf{x}$. Represente con $x_1, x_2, \ldots, x_n$ los valores observados en una muestra. En el caso de una variable aleatoria discreta, la interpretación es muy clara. La cantidad $L(x_1, x_2, \ldots, x_n; \theta)$, *la verosimilitud de la muestra*, es la siguiente probabilidad conjunta:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, \theta),$$

que es la probabilidad de obtener los valores muestrales $x_1, x_2, \ldots, x_n$. Para el caso discreto el estimador de máxima verosimilitud es el que da como resultado un valor máximo para esta probabilidad conjunta, o el que maximiza la probabilidad de la muestra.

Considere un ejemplo ficticio en el cual se inspeccionan tres artículos que salen de una línea de ensamble. Los artículos se clasifican como defectuosos o no defectuosos, de manera que se aplica el proceso de Bernoulli. La inspección de los tres artículos da como resultado dos artículos no defectuosos seguidos por uno defectuoso. Nos interesa estimar p, la proporción de artículos no defectuosos en el proceso. La probabilidad de la muestra para este ejemplo es dada por

$$p \cdot p \cdot q = p^2 q = p^2 - p^3,$$

donde $q = 1 - p$. La estimación de máxima verosimilitud daría un estimado de p para el que se maximiza la verosimilitud. Resulta claro que si diferenciamos la verosimilitud respecto a p, igualamos la derivada a cero y la resolvemos, obtenemos el valor

$$\hat{p} = \frac{2}{3}.$$

Entonces, desde luego, en esta situación $\hat{p} = 2/3$ es la proporción muestral defectuosa y, por ello, un estimador razonable de la probabilidad de un artículo defectuoso. El lector debería intentar comprender que la filosofía de la estimación de máxima verosimilitud proviene de la noción de que el estimador razonable de un parámetro que se basa en información muestral *es el valor del parámetro que produce la mayor probabilidad de obtener la muestra.* Ésta es, de hecho, la interpretación para el caso discreto, ya que la verosimilitud es la probabilidad de observar de manera conjunta los valores en la muestra.

Así, mientras que la interpretación de la función de verosimilitud como una probabilidad conjunta se limita al caso discreto, la noción de máxima verosimilitud se extiende a la estimación de parámetros de una distribución continua. Presentamos ahora una definición formal de la estimación de máxima verosimilitud.

Definición 9.3: Dadas las observaciones independientes $x_1, x_2, \ldots, x_n$ de una función de densidad de probabilidad (caso continuo) o de una función de masa de probabilidad (caso discreto) $f(\mathbf{x}, \theta)$, el estimador de máxima verosimilitud $\hat{\theta}$ es el que maximiza la función de probabilidad

$$L(x_1, x_2, \ldots, x_n; \theta) = f(\mathbf{x}; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta).$$

Muy a menudo conviene trabajar con el logaritmo natural de la función de verosimilitud para encontrar el máximo de esa función. Considere el siguiente ejemplo acerca del parámetro μ de una distribución de Poisson.

Ejemplo 9.20: Considere una distribución de Poisson con la siguiente función de masa de probabilidad

$$f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

Suponga que se toma una muestra aleatoria $x_1, x_2, \ldots, x_n$ de la distribución. ¿Cuál es la estimación de máxima verosimilitud de μ?

***Solución*:** La función de probabilidad es

$$L(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} f(x_i|\mu) = \frac{e^{-n\mu}\mu^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Considere ahora

$$\ln L(x_1, x_2, \ldots, x_n; \mu) = -n\mu + \sum_{i=1}^{n} x_i \ln \mu - \ln \prod_{i=1}^{n} x_i!$$

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \mu)}{\partial \mu} = -n + \sum_{i=1}^{n} \frac{x_i}{\mu}.$$

Resolver para $\hat{\mu}$, el estimador de máxima verosimilitud, implica definir la derivada para cero y resolver para el parámetro. Por consiguiente,

$$\hat{\mu} = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}.$$

La segunda derivada de la función de verosimilitud logarítmica es negativa, lo cual implica que la solución anterior realmente es un máximo. Como μ es la media de la distribución de Poisson (capítulo 5), el promedio muestral en realidad parecería ser un estimador razonable. ∎

El siguiente ejemplo presenta el uso del método de máxima verosimilitud para calcular estimados de dos parámetros. Simplemente encontramos los valores de los parámetros que maximizan (de forma conjunta) la función de probabilidad.

Ejemplo 9.21: Considere una muestra aleatoria $x_1, x_2, \ldots, x_n$ de una distribución normal $N(\mu, \sigma)$. Calcule los estimadores de máxima verosimilitud para μ y σ^2.

***Solución*:** La función de verosimilitud para la distribución normal es

$$L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right].$$

Al usar logaritmos obtenemos

$$\ln L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2.$$

Por lo tanto,

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma^2}\right)$$

y

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Al igualar ambas derivadas a cero, obtenemos

$$\sum_{i=1}^{n} x_i - n\mu = 0 \quad \text{y} \quad n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2.$$

Por consiguiente, el estimador de máxima verosimilitud de μ es dado por

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

que es un resultado satisfactorio, ya que $\bar{x}$ ha desempeñado un papel tan importante en este capítulo como un estimador puntual de μ. Por otro lado, el estimador de máxima verosimilitud de σ^2 es

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Al verificar la matriz derivada parcial de segundo orden se confirma que la solución da como resultado el máximo de la función de verosimilitud. ∎

Resulta interesante notar la distinción entre el estimador de máxima verosimilitud de σ^2 y el estimador insesgado S^2 que se presentó al principio de este capítulo. Los numeradores son idénticos, desde luego, y el denominador lo constituyen los "grados de libertad" $n - 1$ para el estimador insesgado, y n para el estimador de máxima verosimilitud. Los estimadores de máxima verosimilitud no necesariamente gozan de la propiedad de carecer de sesgo. Sin embargo, los estimadores de máxima verosimilitud tienen importantes propiedades asintóticas.

Ejemplo 9.22: Suponga que en un estudio biomédico se utilizan 10 ratas a las que después de inyectarles células cancerosas se les suministra un fármaco contra el cáncer diseñado para aumentar su tasa de supervivencia. Los tiempos de supervivencia, en meses, son 14, 17, 27, 18, 12,

8, 22, 13, 19 y 12. Suponga que se trata de una distribución exponencial. Calcule un estimado de máxima verosimilitud de la supervivencia media.

***Solución*:** Del capítulo 6 sabemos que la función de densidad de probabilidad para la variable aleatoria exponencial X es

$$f(x, \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Por consiguiente, la función de verosimilitud logarítmica de los datos, dado que $n = 10$, es

$$\ln L(x_1, x_2, \ldots, x_{10}; \beta) = -10 \ln \beta - \frac{1}{\beta} \sum_{i=1}^{10} x_i.$$

Si se establece que

$$\frac{\partial \ln L}{\partial \beta} = -\frac{10}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{10} x_i = 0$$

implica que

$$\hat{\beta} = \frac{1}{10} \sum_{i=1}^{10} x_i = \bar{x} = 16.2.$$

Si se evalúa la segunda derivada de la función de verosimilitud logarítmica en el valor $\hat{\beta}$ anterior se produce un valor negativo. Como resultado, el estimador del parámetro β, la media de la población, es el promedio muestral $\bar{x}$. ∎

El siguiente ejemplo ilustra el estimador de máxima verosimilitud para una distribución que no se incluye en los capítulos anteriores.

Ejemplo 9.23: Se sabe que una muestra que consta de los valores 12, 11.2, 13.5, 12.3, 13.8 y 11.9 proviene de una población con la siguiente función de densidad

$$f(x; \theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & x > 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

donde $\theta > 0$. Calcule la estimación de máxima verosimilitud de θ.

***Solución*:** La función de verosimilitud de n observaciones de esta población se escribe como

$$L(x_1, x_2, \ldots, x_{10}; \theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^{\theta+1}} = \frac{\theta^n}{(\prod_{i=1}^{n} x_i)^{\theta+1}},$$

lo cual implica que

$$\ln L(x_1, x_2, \ldots, x_{10}; \theta) = n \ln(\theta) - (\theta + 1) \sum_{i=1}^{n} \ln(x_i).$$

Si establecemos que $0 = \frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \ln(x_i)$ da como resultado

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln(x_i)}$$

$$= \frac{6}{\ln(12) + \ln(11.2) + \ln(13.5) + \ln(12.3) + \ln(13.8) + \ln(11.9)} = 0.3970.$$

Como la segunda derivada de L es $-n/\theta^2$, que siempre es negativa, la función de probabilidad alcanza su valor máximo en $\hat{\theta}$. ■

Comentarios adicionales respecto a la estimación de máxima verosimilitud

Un análisis detallado de las propiedades de la estimación de máxima verosimilitud está fuera del alcance de este libro y, por lo general, es un tema importante en un curso teórico de estadística inferencial. El método de máxima verosimilitud permite al analista utilizar el conocimiento de la distribución para determinar un estimador adecuado. *El método de máxima verosimilitud no se puede aplicar si no se conoce la distribución subyacente.* En el ejemplo 9.21 aprendimos que el estimador de máxima verosimilitud no necesariamente carece de sesgo. El estimador de máxima verosimilitud es insesgado *asintóticamente* o *en el límite*; es decir, la magnitud del sesgo se aproxima a cero a medida que la muestra se hace más grande. Al principio de este capítulo examinamos la noción de eficacia, que se vincula con la propiedad de la varianza de un estimador. Los estimadores de máxima verosimilitud tienen propiedades de varianza deseables en el límite. El lector debería consultar la obra de Lehmann y D'Abrera (1998) para más detalles.

Ejercicios

9.81 Suponga que hay n ensayos $x_1, x_2, \ldots, x_n$ de un proceso de Bernoulli con parámetro p, la probabilidad de un éxito. Esto es, la probabilidad de r éxitos es dada por $\binom{n}{r} p^r (1-p)^{n-r}$. Determine el estimador de máxima verosimilitud para el parámetro p.

9.82 Considere la distribución logarítmica normal con la función de densidad dada en la sección 6.9. Suponga que tiene una muestra aleatoria $x_1, x_2, \ldots, x_n$ de una distribución logarítmica normal.

a) Escriba la función de verosimilitud.
b) Desarrolle los estimadores de máxima verosimilitud de μ y σ^2.

9.83 Considere una muestra aleatoria de $x_1, \ldots, x_n$ obtenida de la distribución gamma descrita en la sección 6.6. Suponga que conoce el parámetro α, el cual digamos que es 5, y con base en esto determine la estimación de máxima verosimilitud para el parámetro β.

9.84 Considere una muestra aleatoria de $x_1, x_2, \ldots, x_n$ observaciones de una distribución de Weibull con parámetros α y β, y la siguiente función de densidad

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0, \\ 0, & \text{en cualquier caso,} \end{cases}$$

para $\alpha, \beta > 0$.

a) Escriba la función de verosimilitud.
b) Escriba las ecuaciones que al resolverse proporcionan los estimadores de máxima verosimilitud de α y β.

9.85 Considere una muestra aleatoria de $x_1, \ldots, x_n$ obtenida de una distribución uniforme $U(0, \theta)$, con el parámetro θ desconocido, donde $\theta > 0$. Determine el estimador de máxima verosimilitud de θ.

9.86 Considere las observaciones independientes de $x_1, x_2, \ldots, x_n$ de la distribución gamma que se analizó en la sección 6.6.

a) Escriba la función de verosimilitud.
b) Escriba un conjunto de ecuaciones que, cuando se resuelven, proporcionan los estimadores de máxima verosimilitud de α y β.

9.87 Considere un experimento hipotético en el que un hombre que tiene un hongo utiliza un medicamento fungicida y se cura. Por lo tanto, considere que se trata de una muestra de una distribución de Bernoulli con la siguiente función de probabilidad

$$f(x) = p^x q^{1-x}, \quad x = 0, 1,$$

donde p es la probabilidad de un éxito (curación) y $q = 1 - p$. Ahora, desde luego, la información muestral da $x = 1$. Escriba un procedimiento que demuestre que $\hat{p} = 1.0$ es el estimador de máxima probabilidad de curación.

9.88 Considere la observación X de la distribución binomial negativa dada en la sección 5.4. Calcule el estimador de máxima verosimilitud para p, suponiendo que se conoce k.

Ejercicios de repaso

9.89 Considere dos estimadores de σ^2 para una muestra $x_1, x_2, \ldots, x_n$ que se extrae de una distribución normal con media μ y varianza σ^2. Los estimadores son el estimador insesgado $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ y el estimador de máxima verosimilitud $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Analice las propiedades de la varianza de estos dos estimadores.

9.90 De acuerdo con el *Roanoke Times*, McDonald's vendió 42.1% de la participación del mercado de hamburguesas. Una muestra aleatoria de 75 hamburguesas vendidas reveló que 28 de ellas fueron vendidas por McDonald's. Utilice el material de la sección 9.10 para determinar si esta información respalda la afirmación del *Roanoke Times*.

9.91 Se afirma que un individuo podrá reducir, en un lapso de 2 semanas, un promedio de 4.5 kilogramos de peso con una nueva dieta. Los pesos de 7 mujeres que siguieron esta dieta se registraron antes y después de un periodo de 2 semanas.

Mujer	Peso antes	Peso después
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	69.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4

Pruebe la afirmación sobre la dieta calculando un intervalo de confianza del 95% para la diferencia media en el peso. Suponga que las diferencias de los pesos se distribuyen de forma aproximadamente normal.

9.92 En Virginia Tech se realizó un estudio para determinar si se puede utilizar el fuego como una herramienta de control viable para aumentar la cantidad de forraje disponible para los venados durante los meses críticos a finales del invierno y principios de la primavera. El calcio es un elemento necesario para las plantas y los animales. La cantidad que la planta toma y almacena está estrechamente correlacionada con la cantidad presente en el suelo. Se formuló la hipótesis de que el fuego podría cambiar los niveles de calcio presentes en el suelo y, por lo tanto, influir en la cantidad disponible para los venados. Se seleccionó una extensión grande de tierra en el bosque Fishburn para provocar un incendio controlado. Justo antes de la quema se tomaron muestras de suelo de 12 parcelas con la misma área y se analizaron para verificar su contenido de calcio. Después del incendio se volvieron a analizar los niveles de calcio en las mismas parcelas. Los valores obtenidos, en kilogramos por parcela, se presentan en la siguiente tabla:

	Nivel de calcio (kg/parcela)	
Parcela	Antes del incendio	Después del incendio
1	50	9
2	50	18
3	82	45
4	64	18
5	82	18
6	73	9
7	77	32
8	54	9
9	23	18
10	45	9
11	36	9
12	54	9

Construya un intervalo de confianza del 95% para la diferencia media en los niveles de calcio presentes en el suelo antes y después del incendio controlado. Suponga que la distribución de las diferencias en los niveles de calcio es aproximadamente normal.

9.93 El dueño de un gimnasio afirma que una persona podrá reducir, en un periodo de 5 días, un promedio de 2 centímetros en su talla de cintura con un nuevo programa de ejercicios. En la siguiente tabla se presentan

las tallas de cintura de 6 hombres que participaron en este programa de ejercicios antes y después del periodo de 5 días:

Hombre	Talla de cintura antes	Talla de cintura después
1	90.4	91.7
2	95.5	93.9
3	98.7	97.4
4	115.9	112.8
5	104.0	101.3
6	85.6	84.0

Mediante el cálculo de un intervalo de confianza del 95% para la reducción media en la talla de cintura determine si la afirmación del dueño del gimnasio es válida. Suponga que la distribución de las diferencias en las tallas de cintura antes y después del programa es aproximadamente normal.

9.94 El Departamento de Ingeniería Civil del Virginia Tech comparó una técnica de ensayo modificada (M-5 hr) para recuperar coliformes fecales en residuos líquidos (charcos) de agua de lluvia en una área urbana con la técnica del número más probable (NMP). El departamento recolectó un total de 12 muestras de tales residuos y las analizó con las dos técnicas. Los conteos de coliformes fecales por 100 mililitros se registraron en la siguiente tabla:

Muestra	Conteo NMP	Conteo con M-5 hr
1	2300	2010
2	1200	930
3	450	400
4	210	436
5	270	4100
6	450	2090
7	154	219
8	179	169
9	192	194
10	230	174
11	340	274
12	194	183

Construya un intervalo de confianza del 90% para la diferencia entre el conteo medio de coliformes fecales que se obtuvo con la técnica M-5 hr y el que se obtuvo con la NMP. Suponga que las diferencias en los conteos se distribuyen de forma aproximadamente normal.

9.95 Se llevó a cabo un experimento para determinar si el acabado superficial tiene un efecto en el límite de resistencia a la fatiga del acero. Una teoría indica que el pulido aumenta el límite medio de resistencia a la fatiga (para la flexión inversa). Desde un punto de vista práctico, el pulido no debería tener efecto alguno sobre la desviación estándar del límite de resistencia a la fatiga, el cual se sabe, a partir de la realización de diversos experimentos de límite de resistencia a la fatiga, que es de 4000 psi. Se realiza un experimento sobre acero al carbono al 0.4% usando especímenes sin pulido y especímenes con pulido suave. Los datos son los siguientes:

Límite de fatiga (psi)	
Acero al carbono al 0.4%	Acero al carbono al 0.4% sin pulir
85,500	82,600
91,900	82,400
89,400	81,700
84,000	79,500
89,900	79,400
78,700	69,800
87,500	79,900
83,100	83,400

Calcule un intervalo de confianza del 95% para la diferencia entre las medias de la población para los dos métodos. Suponga que las poblaciones se distribuyen de forma aproximadamente normal.

9.96 Un antropólogo está interesado en determinar la proporción de individuos de dos tribus indias que tienen doble remolino de cabello en la zona occipital. Suponga que toma muestras independientes de cada una de las dos tribus y encuentra que 24 de 100 individuos de la tribu *A* y 36 de 120 individuos de la tribu *B* poseen tal característica. Construya un intervalo de confianza del 95% para la diferencia $p_B - p_A$ entre las proporciones de estas dos tribus con remolinos de cabello en la zona occipital.

9.97 Un fabricante de planchas eléctricas produce estos artículos en dos plantas en las que las partes pequeñas son surtidas por el mismo proveedor. El fabricante puede ahorrar algo si le compra a un proveedor local los termostatos para la planta *B*. Para probar si estos nuevos termostatos son tan precisos como los anteriores le compra sólo un lote al proveedor local y los prueba en planchas a 550°F. Al final lee con un termopar las temperaturas reales y las redondea al siguiente 0.1°F más cercano. Los datos son los siguientes:

Proveedor nuevo (°F)					
530.3	559.3	549.4	544.0	551.7	566.3
549.9	556.9	536.7	558.8	538.8	543.3
559.1	555.0	538.6	551.1	565.4	554.9
550.0	554.9	554.7	536.1	569.1	
Proveedor anterior (°F)					
559.7	534.7	554.8	545.0	544.6	538.0
550.7	563.1	551.1	553.8	538.8	564.6
554.5	553.0	538.4	548.3	552.9	535.1
555.0	544.8	558.4	548.7	560.3	

Calcule un intervalo de confianza de 95% para σ_1^2/σ_2^2 y para σ_1/σ_2, donde σ_1^2 y σ_2^2 son las varianzas de la

población de las lecturas de los termostatos del proveedor nuevo y del anterior, respectivamente.

9.98 Se afirma que la resistencia del alambre *A* es mayor que la del alambre *B*. Un experimento sobre los alambres muestra los siguientes resultados (en ohms):

Alambre *A*	Alambre *B*
0.140	0.135
0.138	0.140
0.143	0.136
0.142	0.142
0.144	0.138
0.137	0.140

Suponga varianzas iguales y explique a qué conclusiones llega si se basa en esto.

9.99 Una forma alternativa de estimación se lleva a cabo a través del método de momentos. El método consiste en igualar la media y la varianza de la población con las correspondientes media muestral $\bar{x}$ y varianza muestral s^2, y resolver para los parámetros; el resultado son los **estimadores por momentos**. En el caso de un solo parámetro sólo se utilizan las medias. Argumente por qué en el caso de la distribución de Poisson el estimador de máxima verosimilitud y los estimadores por momentos son iguales.

9.100 Especifique los estimadores por momentos para μ y σ^2 para la distribución normal.

9.101 Especifique los estimadores por momentos para μ y σ^2 para la distribución logarítmica normal.

9.102 Especifique los estimadores por momentos para α y β en el caso de la distribución gamma.

9.103 Se realizó una encuesta con el fin de comparar los sueldos de administradores de plantas químicas empleados en dos áreas del país: el norte y el centro-occidente. Se eligió una muestra aleatoria independiente de 300 gerentes de planta para cada una de las dos áreas. A tales gerentes se les preguntó el monto de su sueldo anual. Los resultados fueron los siguientes:

Norte	Centro-Occidente
$\bar{x}_1 = \$102,300$	$\bar{x}_2 = \$98,500$
$s_1 = \$5700$	$s_2 = \$3800$

a) Construya un intervalo de confianza del 99% para $\mu_1 - \mu_2$, la diferencia en los sueldos medios.

b) ¿Qué supuso en el inciso *a*) acerca de la distribución de los sueldos anuales para las dos áreas? ¿Es necesaria la suposición de normalidad? Explique su respuesta.

c) ¿Qué supuso acerca de las dos varianzas? ¿Es razonable la suposición de igualdad de varianzas? ¡Explique!

9.104 Considere el ejercicio de repaso 9.103. Suponga que los datos aún no se han recabado. Suponga también que los estadísticos previos sugieren que $\sigma_1 = \sigma_2 = \$4000$. ¿Los tamaños de las muestras en el ejercicio de repaso 9.103 son suficientes para producir un intervalo de confianza del 95% si $\mu_1 - \mu_2$ tiene una anchura de sólo $1000? Presente el desarrollo completo.

9.105 Un sindicato se preocupa por el notorio ausentismo de sus miembros. Los líderes del sindicato siempre habían afirmado que, en un mes típico, el 95% de sus afiliados estaban ausentes menos de 10 horas al mes. El sindicato decide verificar esto revisando una muestra aleatoria de 300 de sus miembros. Se registra el número de horas de ausencia para cada uno de los 300 miembros. Los resultados son $\bar{x} = 6.5$ horas y $s = 2.5$ horas. Utilice los datos para responder esa afirmación utilizando un límite de tolerancia unilateral y eligiendo un nivel de confianza del 99%. Asegúrese de aplicar lo que ya sabe acerca del cálculo del límite de tolerancia.

9.106 Se seleccionó una muestra aleatoria de 30 empresas que comercializan productos inalámbricos para determinar la proporción de tales empresas que implementaron software nuevo para aumentar la productividad. Resultó que 8 de las 30 empresas habían implementado tal software. Calcule un intervalo de confianza del 95% en p, la proporción verdadera de ese tipo de empresas que implementaron el nuevo software.

9.107 Remítase al ejercicio de repaso 9.106. Suponga que se desea saber si la estimación puntual $\hat{p} = 8/30$ es lo suficientemente precisa porque el intervalo de confianza alrededor de p no es tan estrecho como se requiere. Utilice $\hat{p}$ como el estimado de p para determinar cuántas empresas habría que incluir en una muestra para obtener un intervalo de confianza del 95% con una anchura de sólo 0.05.

9.108 Un fabricante produce un artículo que se clasifica como "defectuoso" o "no defectuoso". Para estimar la proporción de productos defectuosos se tomó una muestra aleatoria de 100 artículos de la producción y se encontraron 10 defectuosos. Después de aplicar un programa de mejoramiento de la calidad se volvió a realizar el experimento. Se tomó una nueva muestra de 100 artículos y esta vez sólo 6 salieron defectuosos.

a) Dado un intervalo de confianza del 95% de $p_1 - p_2$, donde p_1 y p_2 representan la proporción de artículos defectuosos de la población antes y después del mejoramiento, respectivamente.

b) ¿Hay información en el intervalo de confianza que se encontró en el inciso *a*) que sugiera que $p_1 > p_2$? Explique su respuesta.

9.109 Se utiliza una máquina para llenar cajas de un producto en una operación de la línea de ensamble. Gran parte del interés se centra en la variabilidad del número de onzas del producto en la caja. Se sabe que la desviación estándar en el peso del producto es de 0.3 onzas. Se realizan mejoras y luego se toma una muestra aleatoria de 20 cajas, y se encuentra que la varianza de la muestra es de 0.045 onzas^2. Calcule un intervalo de confianza del 95% de la varianza del peso del producto. Si considera el rango del intervalo de confianza, ¿le parece que el mejoramiento en el proceso incrementó la calidad en lo que se refiere a la variabilidad? Suponga normalidad en la distribución del peso del producto.

9.110 Un grupo de consumidores está interesado en comparar los costos de operación de dos diferentes tipos de motor para automóvil. El grupo encuentra 15 propietarios cuyos automóviles tienen motor tipo *A* y 15 que tienen motor tipo *B*. Los 30 propietarios compraron sus automóviles más o menos al mismo tiempo y todos llevaron buenos registros en cierto periodo de 12 meses. Los consumidores encontraron, además, que los propietarios recorrieron aproximadamente el mismo número de millas. Los estadísticos de costo son $\bar{y}_A = \$87.00/1000$ millas, $\bar{y}_B = \$75.00/1000$ millas, $s_A = \$5.99$ y $s_B = \$4.85$. Calcule un intervalo de confianza del 95% para estimar $\mu_A - \mu_B$, la diferencia en el costo medio de operación. Suponga normalidad y varianzas iguales.

9.111 Considere el estadístico S_p^2, el estimado agrupado de σ^2 que se estudió en la sección 9.8 y que se utiliza cuando se está dispuesto a suponer que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Demuestre que el estimador es insesgado para σ^2 [es decir, demuestre que $E(S_p^2) = \sigma^2$]. Puede utilizar los resultados de cualquier teorema o ejemplo de este capítulo.

9.112 Un grupo de investigadores del factor humano están interesados en saber cómo reaccionan los pilotos aviadores ante un estímulo dispuesto de cierta manera en la cabina del avión. Para lograr su objetivo realizaron un experimento de simulación en un laboratorio, el cual incluyó a 15 pilotos, los que presentaron un tiempo de reacción promedio de 3.2 segundos y una desviación estándar muestral de 0.6 segundos. Resulta de interés caracterizar el extremo, es decir, el escenario del peor caso. Para conseguir esto realice lo siguiente:

a) Determine un importante límite de confianza unilateral específico del 99% del tiempo medio de reacción. ¿Qué suposición, si la hubiera, debería hacer acerca de la distribución de los tiempos de reacción?

b) Determine un intervalo unilateral de predicción del 99% e interprete su significado. ¿Debería usted suponer algo sobre la distribución de los tiempos de reacción para calcular este límite?

c) Calcule un límite de tolerancia unilateral con una confianza del 99% que incluya al 95% de los tiempos de reacción. Nuevamente, de ser necesario, interprete o suponga algo acerca de la distribución. [Nota: Los valores del límite de tolerancia unilateral también se incluyen en la tabla A.7].

9.113 Cierto proveedor fabrica un tipo de tapete de hule que vende a las empresas automotrices. El material que utiliza para los tapetes debe tener ciertas características de dureza. Ocasionalmente detecta tapetes defectuosos en el proceso y los rechaza. El proveedor afirma que la proporción de tapetes defectuosos es de 0.05, pero como un cliente que compró los tapetes desafió su afirmación, realizó un experimento en el que se probaron 400 tapetes y se encontraron 17 defectuosos.

a) Calcule un intervalo de confianza bilateral del 95% de la proporción de tapetes defectuosos.

b) Calcule un intervalo de confianza unilateral del 95% adecuado de la proporción de tapetes defectuosos.

c) Interprete los intervalos de ambos incisos y comente acerca de la afirmación hecha por el proveedor.

9.15 Posibles riesgos y errores conceptuales: relación con el material de otros capítulos

El concepto de *intervalo de confianza de muestra grande* en una población a menudo confunde a los alumnos principiantes. Se basa en la idea de que incluso cuando se desconoce σ y no se está convencido de que la distribución que se muestrea es normal, se puede calcular un intervalo de confianza para μ a partir de

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

En la práctica es común que se utilice esta fórmula cuando la muestra es demasiado pequeña. El origen de este intervalo de muestra grande es, por supuesto, el teorema del

límite central (TLC), con el cual la normalidad no es necesaria. Aquí el TLC requiere una σ conocida, de la cual s sólo es un estimado. Por lo tanto, n debe ser al menos tan grande como 30 y la distribución subyacente debe tener una simetría similar, en cuyo caso el intervalo sigue siendo una aproximación.

Hay casos en que la aplicación práctica del material de este capítulo depende en gran medida del contexto específico. Un ejemplo muy importante es el uso de la distribución t para el intervalo de confianza de μ cuando se desconoce σ. En términos estrictos, el uso de la distribución t requiere que la distribución de donde se toma la muestra sea normal. Sin embargo, es bien sabido que cualquier aplicación de la distribución t es razonablemente insensible, es decir, **robusta**, a la suposición de normalidad. Esto representa una de esas situaciones afortunadas que ocurren con frecuencia en el campo de la estadística, donde no se sostiene un supuesto básico y "¡todo resulta bien!" Sin embargo, la población de la que se toma la muestra no se puede desviar mucho de la normalidad. Por consiguiente, a menudo se recurrirá a las gráficas de probabilidad normal estudiadas en el capítulo 8 y las pruebas de bondad del ajuste que se presentarán en el capítulo 10 para atribuir algún sentido de "cercanía a la normalidad". Esta idea de "robustez a la normalidad" se volverá a presentar en el capítulo 10.

Por experiencia sabemos que uno de los más graves "usos incorrectos de la estadística" en la práctica surge de la confusión sobre las diferencias en la interpretación de los tipos de intervalos estadísticos. Por consiguiente, la subsección de este capítulo en la que se examinan las diferencias entre los tres tipos de intervalos es importante. Es muy probable que en la práctica **se utilice en exceso el intervalo de confianza**, es decir, que se emplee cuando no es la media lo que interesa en realidad, sino la cuestión de: "¿en dónde va a caer la siguiente observación?", o la a menudo más importante cuestión de: "¿en dónde se ubica la mayor parte de la distribución?" Éstas son preguntas fundamentales que no se pueden responder calculando un intervalo de la media. A menudo resulta confusa la interpretación de un intervalo de confianza. Es tentador concluir que hay una probabilidad de 0.95 de que el parámetro caiga dentro del intervalo. Aunque se trata de una interpretación correcta del **intervalo posterior bayesiano** (para mayores referencias sobre la inferencia bayesiana véase el capítulo 18), no es una interpretación adecuada de la frecuencia.

El intervalo de confianza tan sólo sugiere que si se realiza el experimento y los datos se observan una y otra vez, aproximadamente 95% de tales intervalos contendrá el parámetro verdadero. Cualquier alumno principiante de la estadística práctica debería tener muy claras las diferencias entre estos intervalos estadísticos.

Otro posible y grave uso incorrecto de la estadística es el que se cometería si se aplicara la distribución χ^2 a un intervalo de confianza de una sola varianza. De nuevo, se supone normalidad en la distribución de donde se toma la muestra. A diferencia del resultado de utilizar la distribución t, la prueba χ^2 para esta aplicación **no es robusta para la suposición de normalidad** (esto significa que cuando la distribución subyacente no es normal, la distribución muestral de $\frac{(n-1)S^2}{\sigma^2}$ se aparta mucho de χ^2). En consecuencia, el uso estricto de la prueba de bondad de ajuste (véase el capítulo 10) y de las gráficas de probabilidad normal, o de la prueba y las gráficas, puede ser muy importante en esos contextos. En los siguientes capítulos se proporcionará más información sobre este tema general.

Capítulo 10

Pruebas de hipótesis de una y dos muestras

10.1 Hipótesis estadísticas: conceptos generales

Como se expuso en el capítulo 9, a menudo el problema al que se enfrentan el científico o el ingeniero no es tanto la estimación de un parámetro de la población, sino la formación de un procedimiento de decisión que se base en los datos y que pueda producir una conclusión acerca de algún sistema científico. Por ejemplo, un investigador médico puede decidir con base en evidencia experimental si beber café incrementa el riesgo de cáncer en los seres humanos; un ingeniero quizá tenga que decidir con base en datos muestrales si hay una diferencia entre la precisión de un tipo de medidor y la de otro; o tal vez un sociólogo desee reunir los datos apropiados que le permitan decidir si el tipo de sangre y el color de ojos de un individuo son variables independientes. En cada uno de estos casos el científico o el ingeniero *postulan* o *conjeturan* algo acerca de un sistema. Además, cada uno debe utilizar datos experimentales y tomar decisiones basadas en ellos. En cada caso la conjetura se puede expresar en forma de hipótesis estadística. Los procedimientos que conducen a la aceptación o al rechazo de hipótesis estadísticas como éstas comprenden una área importante de la inferencia estadística. Empecemos por definir con precisión lo que entendemos por **hipótesis estadística**.

Definición 10.1: Una **hipótesis estadística** es una aseveración o conjetura respecto a una o más poblaciones.

La verdad o falsedad de una hipótesis estadística nunca se sabe con absoluta certeza, a menos que se examine toda la población, lo cual, por supuesto, sería poco práctico en la mayoría de las situaciones. En vez de eso se toma una muestra aleatoria de la población de interés y se utilizan los datos contenidos en ella para proporcionar evidencia que respalde o no la hipótesis. La evidencia de la muestra que es inconsistente con la hipótesis planteada conduce al rechazo de la misma.

El papel que desempeña la probabilidad en la prueba de hipótesis

Debería quedar claro al lector que un procedimiento de toma de decisiones debe implicar la conciencia de la *probabilidad de llegar a una conclusión errónea*. Por ejemplo, suponga que la hipótesis que postuló el ingeniero es que la fracción p de artículos defectuosos en cierto proceso es 0.10. El experimento consiste en observar una muestra aleatoria del producto en cuestión. Suponga que se prueban 100 artículos y que se encuentran 12 defectuosos. Es razonable concluir que esta evidencia no rechaza la condición de que el parámetro binomial $p = 0.10$, por lo que puede provocar que no se rechace la hipótesis. Sin embargo, también puede provocar que no se refute $p = 0.12$, o quizá incluso $p = 0.15$. Como resultado, el lector se debe acostumbrar a la idea de que **el rechazo de una hipótesis implica que fue refutada por la evidencia de la muestra**. En otras palabras, **el rechazo significa que existe una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera**. Por ejemplo, en la hipótesis de la proporción de artículos defectuosos, una muestra de 100 artículos que revela que hay 20 defectuosos es ciertamente evidencia para el rechazo. ¿Por qué? Si en realidad $p = 0.10$, la probabilidad de obtener 20 o más artículos defectuosos es aproximadamente de 0.002. Con el pequeño riesgo resultante de llegar a una conclusión errónea parecería seguro **rechazar la hipótesis** de que $p = 0.10$. En otras palabras, el rechazo de una hipótesis tiende a casi "descartar" la hipótesis. Por otro lado, es muy importante enfatizar que la aceptación o, más bien, la falta de rechazo no descarta otras posibilidades. Como resultado, *el analista de datos establece una conclusión firme cuando se rechaza una hipótesis*.

En el planteamiento formal de una hipótesis a menudo influye la estructura de la probabilidad de una conclusión errónea. Si el científico está interesado en *apoyar firmemente* un argumento, espera llegar a éste en la forma del rechazo de una hipótesis. Si el investigador médico desea mostrar evidencia sólida a favor del argumento de que beber café aumenta el riesgo de contraer cáncer, la hipótesis a probar debería tener la forma "el riesgo de desarrollar cáncer no aumenta como consecuencia de beber café". Como resultado, el argumento se obtiene mediante un rechazo. De manera similar, para apoyar la afirmación de que un tipo de medidores es más preciso que otro, el ingeniero prueba la hipótesis de que no hay diferencia en la precisión de los dos tipos de medidores.

Lo anterior implica que cuando el analista de datos formaliza la evidencia experimental con base en la prueba de hipótesis, es muy importante el **planteamiento formal de la hipótesis**.

La hipótesis nula y la hipótesis alternativa

La estructura de la prueba de hipótesis se establece usando el término **hipótesis nula**, el cual se refiere a cualquier hipótesis que se desea probar y se denota con H_0. El rechazo de H_0 conduce a la aceptación de una **hipótesis alternativa**, que se denota con H_1. La comprensión de las diferentes funciones que desempeñan la hipótesis nula (H_0) y la hipótesis alternativa (H_1) es fundamental para entender los principios de la prueba de hipótesis. La hipótesis alternativa H_1 por lo general representa la *pregunta que se responderá o la teoría que se probará*, por lo que su especificación es muy importante. La hipótesis nula H_0 *anula o se opone a* H_1 y a menudo es el complemento lógico de H_1. A medida que el lector aprenda más sobre la prueba de hipótesis notará que el analista llega a una de las siguientes dos conclusiones:

rechazar H_0 a favor de H_1 debido a evidencia suficiente en los datos o
no rechazar H_0 debido a evidencia insuficiente en los datos.

Observe que las *conclusiones no implican una "aceptación de* H_0*" formal y literal.* La aseveración de H_0 a menudo representa el "status quo" contrario a una nueva idea, conjetura, etcétera, enunciada en H_1; en tanto que no rechazar H_0 representa la conclusión adecuada. En nuestro ejemplo binomial la cuestión práctica podría ser el interés en que la probabilidad histórica de artículos defectuosos de 0.10 ya no sea verdadera. De hecho, la conjetura podría ser que p excede a 0.10. Entonces podríamos afirmar que

$$H_0: p = 0.10,$$
$$H_1: p > 0.10.$$

Ahora, 12 artículos defectuosos de cada 100 no refutan $p = 0.10$, por lo que la conclusión es "no rechazar H_0". Sin embargo, si los datos revelan 20 artículos defectuosos de cada 100, la conclusión sería "rechazar H_0" a favor de H_1: $p > 0.10$.

Aunque las aplicaciones de la prueba de hipótesis son muy abundantes en trabajos científicos y de ingeniería, quizás el mejor ejemplo para un principiante sea el dilema que enfrenta el jurado en un juicio. Las hipótesis nula y alternativa son

H_0: el acusado es inocente,
H_1: el acusado es culpable.

La acusación proviene de una sospecha de culpabilidad. La hipótesis H_0 (el status quo) se establece en oposición a H_1 y se mantiene a menos que se respalde H_1 con evidencia "más allá de una duda razonable". Sin embargo, en este caso "no rechazar H_0" no implica inocencia, sino sólo que la evidencia fue insuficiente para lograr una condena. Por lo tanto, el jurado no necesariamente *acepta* H_0 sino que *no rechaza* H_0.

10.2 Prueba de una hipótesis estadística

Para ilustrar los conceptos que se utilizan al probar una hipótesis estadística acerca de una población considere el siguiente ejemplo. Se sabe que, después de un periodo de dos años, cierto tipo de vacuna contra un virus que produce resfriado ya sólo es 25% eficaz. Suponga que se eligen 20 personas al azar y se les aplica una vacuna nueva, un poco más costosa, para determinar si protege contra el mismo virus durante un periodo más largo. (En un estudio real de este tipo el número de participantes que reciben la nueva vacuna podría ascender a varios miles. Aquí la muestra es de 20 sólo porque lo único que se busca es demostrar los pasos básicos para realizar una prueba estadística). Si más de 8 individuos de los que reciben la nueva vacuna superan el lapso de 2 años sin contraer el virus, la nueva vacuna se considerará superior a la que se usa en la actualidad. El requisito de que el número exceda a 8 es algo arbitrario, aunque parece razonable, ya que representa una mejoría modesta sobre las 5 personas que se esperaría recibieran protección si fueran inoculadas con la vacuna que actualmente está en uso. En esencia probamos la hipótesis nula de que la nueva vacuna es igual de eficaz después de un periodo de 2 años que la que se utiliza en la actualidad. La hipótesis alternativa es que la nueva vacuna es

mejor, y esto equivale a poner a prueba la hipótesis de que el parámetro binomial para la probabilidad de un éxito en un ensayo dado es $p = ¼$, contra la alternativa de que $p > ¼$. Esto por lo general se escribe como se indica a continuación:

$$H_0: p = 0.25,$$
$$H_1: p > 0.25.$$

El estadístico de prueba

El **estadístico de prueba** en el cual se basa nuestra decisión es X, el número de individuos en nuestro grupo de prueba que reciben protección de la nueva vacuna durante un periodo de al menos 2 años. Los valores posibles de X, de 0 a 20, se dividen en dos grupos: los números menores o iguales que 8 y aquellos mayores que 8. Todos los posibles valores mayores que 8 constituyen la **región crítica**. El último número que observamos al pasar a la región crítica se llama **valor crítico**. En nuestro ejemplo el valor crítico es el número 8. Por lo tanto, si $x > 8$, rechazamos H_0 a favor de la hipótesis alternativa H_1. Si $x \leq 8$, no rechazamos H_0. Este criterio de decisión se ilustra en la figura 10.1.

Figura 10.1: Criterio de decisión para probar $p = 0.25$ contra $p > 0.25$.

La probabilidad de un error tipo 1

El procedimiento de toma de decisiones recién descrito podría conducir a cualquiera de dos conclusiones erróneas. Por ejemplo, es probable que la nueva vacuna no sea mejor que la que se usa en la actualidad (H_0 verdadera) y, sin embargo, en este grupo específico de individuos seleccionados aleatoriamente más de 8 pasan el periodo de 2 años sin contraer el virus. Si rechazáramos H_0 a favor de H_1 cuando, de hecho, H_0 es verdadera, cometeríamos un error que se conoce como **error tipo I**.

Definición 10.2: El rechazo de la hipótesis nula cuando es verdadera se denomina **error tipo I**.

Si 8 o menos miembros del grupo superan exitosamente el periodo de 2 años y no concluimos que la nueva vacuna es mejor cuando en realidad sí lo es (H_1 verdadera), cometemos un segundo tipo de error, el de no rechazar la hipótesis H_0 cuando en realidad es falsa. A este error se le conoce como **error tipo II**.

Definición 10.3: No rechazar la hipótesis nula cuando es falsa se denomina **error tipo II**.

Al probar cualquier hipótesis estadística, hay cuatro situaciones posibles que determinan si nuestra decisión es correcta o errónea. Estas cuatro situaciones se resumen en

la tabla 10.1.

Tabla 10.1: Situaciones posibles al probar una hipótesis estadística.

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

La probabilidad de cometer un error tipo I, también llamada **nivel de significancia**, se denota con la letra griega α. En nuestro ejemplo un error tipo I ocurriría si más de 8 individuos inoculados con la nueva vacuna superan el periodo de 2 años sin contraer el virus y los investigadores concluyen que la nueva vacuna es mejor, cuando en realidad es igual a la vacuna que se utiliza en la actualidad. Por lo tanto, si X es el número de individuos que permanecen sin contraer el virus por al menos dos años,

$$\alpha = P(\text{error tipo I}) = P\left(X > 8 \text{ cuando } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$

$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

Decimos que la hipótesis nula, $p = 1/4$, se prueba al nivel de significancia $\alpha = 0.0409$. En ocasiones el nivel de significancia se conoce como **tamaño de la prueba**. Una región crítica de tamaño 0.0409 es muy pequeña y, por lo tanto, es poco probable que se cometa un error de tipo I. En consecuencia, sería poco probable que más de 8 individuos permanecieran inmunes a un virus durante 2 años utilizando una vacuna nueva que en esencia es equivalente a la que actualmente está en el mercado.

La probabilidad de un error tipo II

La probabilidad de cometer un error tipo II, que se denota con β, es imposible de calcular a menos que tengamos una hipótesis alternativa específica. Si probamos la hipótesis nula $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, entonces podremos calcular la probabilidad de no rechazar H_0 cuando es falsa. Simplemente calculamos la probabilidad de obtener 8 o menos en el grupo que supera el periodo de 2 años cuando $p = 1/2$. En este caso,

$$\beta = P\ (\text{error tipo II}) = P\ \left(X \leq 8 \text{ cuando } p = \frac{1}{2}\right)$$

$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

Se trata de una probabilidad elevada que indica un procedimiento de prueba en el cual es muy probable que se rechace la nueva vacuna cuando, de hecho, es mejor a la que está actualmente en uso. De manera ideal, es preferible utilizar un procedimiento de prueba con el cual haya pocas probabilidades de cometer el error tipo I y el error tipo II.

Es posible que el director del programa de prueba esté dispuesto a cometer un error tipo II si la vacuna más costosa no es significativamente mejor. De hecho, la única

ocasión en la que desea evitar un error tipo II es cuando el verdadero valor de p es de al menos 0.7. Si $p = 0.7$, este procedimiento de prueba da

$$\beta = P(\text{error tipo II}) = P(X \leq 8 \text{ cuando } p = 0.7)$$
$$= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.$$

Con una probabilidad tan pequeña de cometer un error tipo II es muy improbable que se rechace la nueva vacuna cuando tiene una efectividad de 70% después de un periodo de 2 años. A medida que la hipótesis alternativa se aproxima a la unidad, el valor de β tiende a disminuir hasta cero.

El papel que desempeñan α, β y el tamaño de la muestra

Supongamos que el director del programa de prueba no está dispuesto a cometer un error tipo II cuando la hipótesis alternativa $p = 1/2$ es verdadera, aun cuando se encuentre que la probabilidad de tal error es $\beta = 0.2517$. Siempre es posible reducir β aumentando el tamaño de la región crítica. Por ejemplo, considere lo que les sucede a los valores de α y β cuando cambiamos nuestro valor crítico a 7, de manera que todos los valores mayores que 7 caigan en la región crítica y aquellos menores o iguales que 7 caigan en la región de no rechazo. Así, al probar $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, encontramos que

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018$$

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

Al adoptar un nuevo procedimiento de toma de decisiones, reducimos la probabilidad de cometer un error tipo II a costa de aumentar la probabilidad de cometer un error tipo I. Para un tamaño muestral fijo, una disminución en la probabilidad de un error por lo general tendrá como resultado un incremento en la probabilidad del otro error. Por fortuna, **la probabilidad de cometer ambos tipos de errores se puede reducir aumentando el tamaño de la muestra**. Considere el mismo problema usando una muestra aleatoria de 100 individuos. Si más de 36 miembros del grupo superan el periodo de 2 años, rechazamos la hipótesis nula de $p = 1/4$ y aceptamos la hipótesis alternativa de $p > 1/4$. El valor crítico ahora es 36. Todos los valores posibles mayores de 36 constituyen la región crítica y todos los valores posibles menores o iguales que 36 caen en la región de aceptación.

Para determinar la probabilidad de cometer un error tipo I debemos utilizar la aproximación a la curva normal con

$$\mu = np = (100)\left(\frac{1}{4}\right) = 2 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Con respecto a la figura 10.2, necesitamos el área bajo la curva normal a la derecha de $x = 36.5$. El valor z correspondiente es

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

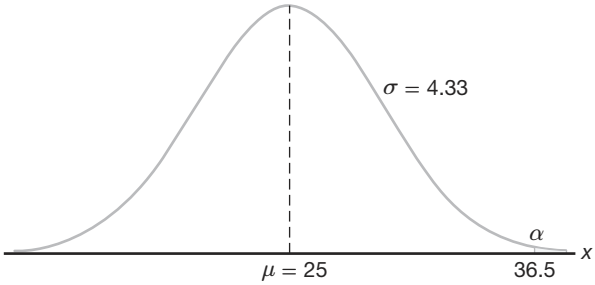


Figura 10.2: Probabilidad de un error tipo I.

En la tabla A.3 encontramos que

$$\alpha = P(\text{error tipo I}) = P\left(X > 36 \text{ cuando } p = \frac{1}{4}\right) \approx P(Z > 2.66)$$
$$= 1 - P(Z < 2.66) = 1 - 0.9961 = 0.0039.$$

Si H_0 es falsa y el verdadero valor de H_1 es $p = 1/2$, determinamos la probabilidad de un error tipo II usando la aproximación a la curva normal con

$$\mu = np = (100)(1/2) = 50 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/2)(1/2)} = 5\ .$$

La probabilidad de que un valor caiga en la región de no rechazo cuando H_0 es verdadera es dada por el área de la región sombreada a la izquierda de $x = 36.5$ en la figura 10.3. El valor z que corresponde a $x = 36.5$ es

$$z = \frac{36.5 - 50}{5} = -2.7.$$

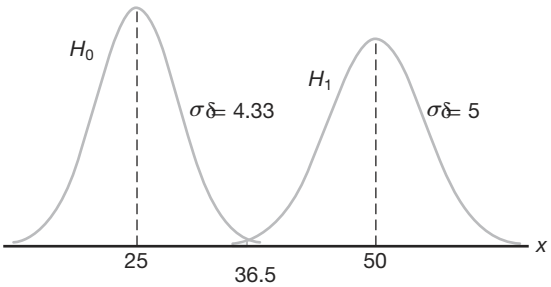


Figura 10.3: Probabilidad de un error tipo II.

Por lo tanto,

$$\beta = P(\text{error tipo II}) = P\left(X \leq 36 \text{ cuando } p = \frac{1}{2}\right) \approx P(Z < -2.7) = 0.0035.$$

Evidentemente, los errores tipo I y tipo II rara vez ocurren si el experimento consta de 100 individuos.

El ejemplo anterior destaca la estrategia del científico en la prueba de hipótesis. Después de que se plantean las hipótesis nula y alternativa es importante considerar la sensibilidad del procedimiento de prueba. Con esto queremos decir que debería determinarse un valor razonable a una α fija para la probabilidad de aceptar de manera errónea H_0, es decir, el valor de β, cuando la verdadera situación representa alguna *desviación importante de* H_0. Por lo general, es posible determinar un valor para el tamaño de la muestra, para el que existe un equilibrio razonable entre los valores de α y β que se calcula de esta manera. El problema de la vacuna es un ejemplo.

Ilustración con una variable aleatoria continua

Los conceptos que se analizan aquí para una población discreta también se pueden aplicar a variables aleatorias continuas. Considere la hipótesis nula de que el peso promedio de estudiantes hombres en cierta universidad es de 68 kilogramos, contra la hipótesis alternativa de que es diferente a 68. Es decir, deseamos probar

$$H_0: \mu = 68,$$
$$H_1: \mu \neq 68.$$

La hipótesis alternativa nos permite la posibilidad de que $\mu < 68$ o $\mu > 68$.

Una media muestral que caiga cerca del valor hipotético de 68 se consideraría como evidencia a favor de H_0. Por otro lado, una media muestral considerablemente menor que o mayor que 68 sería evidencia en contra de H_0 y, por lo tanto, favorecería a H_1. La media muestral es el estadístico de prueba en este caso. Una región crítica para el estadístico de prueba se puede elegir de manera arbitraria como los dos intervalos $\bar{x} < 67$ y $\bar{x} > 69$. La región de no rechazo será entonces el intervalo $67 \leq \bar{x} \leq 69$. Este criterio de decisión se ilustra en la figura 10.4.

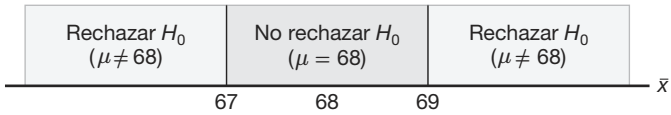


Figura 10.4: Región crítica (en azul).

Utilicemos ahora el criterio de decisión de la figura 10.4 para calcular las probabilidades de cometer los errores tipo I y tipo II cuando probemos la hipótesis nula $\mu = 68$ kilogramos contra la alternativa $\mu \neq 68$ kilogramos.

Suponga que la desviación estándar de la población de pesos es $\sigma = 3.6$. Para muestras grandes podemos sustituir s por σ si no disponemos de ninguna otra estimación de σ. Nuestro estadístico de decisión, que se basa en una muestra aleatoria de tamaño $n = 36$, será $\bar{X}$, el estimador más eficaz de μ. Del teorema del límite central sabemos que la distribución muestral de $\bar{X}$ es aproximadamente normal con desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$.

La probabilidad de cometer un error tipo I, o el nivel de significancia de nuestra prueba, es igual a la suma de las áreas sombreadas en cada cola de la distribución en la figura 10.5. Por lo tanto,

$$\alpha = P(\bar{X} < 67 \text{ cuando } \mu = 68) + P(\bar{X} > 69 \text{ cuando } \mu = 68).$$

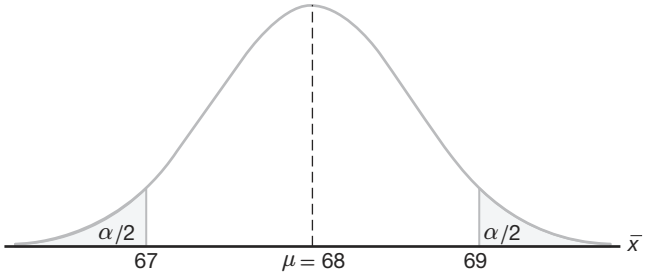


Figura 10.5: Región crítica para probar $\mu = 68$ contra $\mu \neq 68$.

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_0 es verdadera son

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Por lo tanto,

$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

Por consiguiente, 9.5% de todas las muestras de tamaño 36 nos conducirían a rechazar $\mu = 68$ kilogramos cuando, de hecho, ésta es verdadera. Para reducir α tenemos que elegir entre aumentar el tamaño de la muestra o ampliar la región de no rechazo. Suponga que aumentamos el tamaño de la muestra a $n = 64$. Entonces $\sigma_{\bar{X}} = 3.6/8 = 0.45$. En consecuencia,

$$z_1 = \frac{67 - 68}{0.45} = -2.22 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.45} = 2.22.$$

Por lo tanto,

$$\alpha = P(Z < -2.22) + P(Z > 2.22) = 2\,P(Z < -2.22) = 0\,.0264.$$

La reducción de α no es suficiente por sí misma para garantizar un buen procedimiento de prueba. Debemos evaluar β para varias hipótesis alternativas. Si es importante rechazar H_0 cuando la media verdadera sea algún valor $\mu \geq 70$ o $\mu \leq 66$, entonces se debería calcular y examinar la probabilidad de cometer un error tipo II para las alternativas $\mu = 66$ y $\mu = 70$. Debido a la simetría, sólo es necesario considerar la probabilidad de no rechazar la hipótesis nula $\mu = 68$ cuando la alternativa $\mu = 70$ es verdadera. Cuando la media muestral $\bar{x}$ caiga entre 67 y 69, cuando H_1 sea verdadera, resultará un error tipo II. Por lo tanto, remitiéndonos a la figura 10.6 encontramos que

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 70).$$

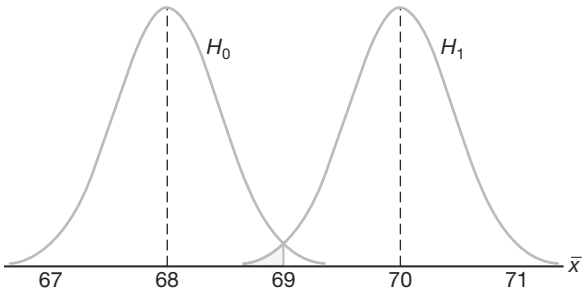


Figura 10.6: Probabilidad del error tipo II al probar $\mu = 68$ contra $\mu = 70$.

Los valores z que corresponden a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_1 es verdadera son

$$z_1 = \frac{67 - 70}{0.45} = -6.67 \quad \text{y} \quad z_2 = \frac{69 - 70}{0.45} = -2.22.$$

Por lo tanto,

$$\begin{aligned}\beta = P(-6.67 < Z < -2.22) &= P(Z < -2.22) - P(Z < -6.67) \\ &= 0.0132 - 0.0000 = 0.0132.\end{aligned}$$

Si el valor verdadero de μ es la alternativa $\mu = 66$, el valor de β nuevamente será 0.0132. Para todos los valores posibles de $\mu < 66$ o $\mu > 70$, el valor de β será incluso más pequeño cuando $n = 64$ y, en consecuencia, habrá poca oportunidad de no rechazar H_0 cuando sea falsa.

La probabilidad de cometer un error tipo II aumenta rápidamente cuando el valor verdadero de μ se aproxima al valor hipotético pero no es igual a éste. Desde luego, ésta suele ser la situación en la que no nos importa cometer un error tipo II. Por ejemplo, si la hipótesis alternativa $\mu = 68.5$ es verdadera, no nos importa cometer un error tipo II al concluir que la respuesta verdadera es $\mu = 68$. La probabilidad de cometer tal error será elevada cuando $n = 64$. Al remitirnos a la figura 10.7, tenemos

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 68.5).$$

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando $\mu = 68.5$ son

$$z_1 = \frac{67 - 68.5}{0.45} = -3.33 \quad \text{y} \quad z_2 = \frac{69 - 68.5}{0.45} = 1.11.$$

Por lo tanto,

$$\begin{aligned}\beta = P(-3.33 < Z < 1.11) &= P(Z < 1.11) - P(Z < -3.33) \\ &= 0.8665 - 0.0004 = 0.8661.\end{aligned}$$

Los ejemplos anteriores ilustran las siguientes propiedades importantes:

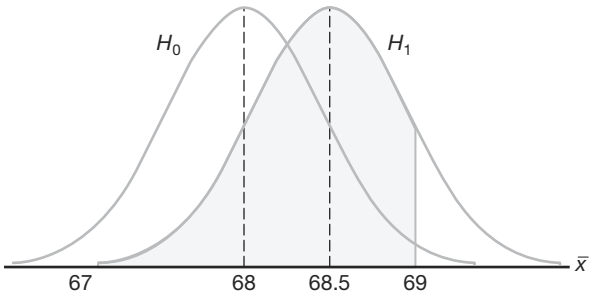


Figura 10.7: Error tipo II para la prueba de $\mu = 68$ contra $\mu = 68.5$.

Propiedades importantes de una prueba de hipótesis

1. Los errores tipo I y tipo II están relacionados. Por lo general una disminución en la probabilidad de cometer uno da como resultado un incremento en la probabilidad de cometer el otro.
2. El tamaño de la región crítica y, por lo tanto, la probabilidad de cometer un error tipo I, siempre se puede reducir ajustando el (los) valor(es) crítico(s).
3. Un aumento en el tamaño de la muestra n reducirá α y β de forma simultánea.
4. Si la hipótesis nula es falsa, β es un máximo cuando el valor verdadero de un parámetro se aproxima al valor hipotético. Cuanto más grande sea la distancia entre el valor verdadero y el valor hipotético, más pequeña será β.

Definición 10.4: La **potencia** de una prueba es la probabilidad de rechazar H_0 dado que una alternativa específica es verdadera.

La potencia de una prueba se puede calcular como $1 - \beta$. A menudo **diferentes tipos de pruebas se comparan contrastando propiedades de potencia**. Considere el caso anterior en el que probamos H_0: $\mu = 68$ y H_1: $\mu \neq 68$. Como antes, suponga que nos interesa evaluar la sensibilidad de la prueba, la cual es determinada por la regla de que no rechazamos H_0 si $67 \leq \bar{x} \leq 69$. Buscamos la capacidad de la prueba para rechazar H_0 de manera adecuada cuando en realidad $\mu = 68.5$. Vimos que la probabilidad de un error tipo II es dada por $\beta = 0.8661$. Por consiguiente, la **potencia** de la prueba es $1 - 0.8661 = 0.1339$. En cierto sentido, la potencia es una medida más sucinta de cuán sensible es la prueba para detectar diferencias entre una media de 68 y otra de 68.5. En este caso, si μ es verdaderamente 68.5, la prueba como se describe *rechazará de forma adecuada H_0 sólo 13.39% de las veces.* Como resultado, la prueba no sería buena si es importante que el analista tenga una oportunidad razonable de distinguir realmente entre una media de 68.0 (que especifica H_0) y una media de 68.5. De lo anterior resulta claro que para producir una potencia deseable, digamos, mayor que 0.8, es necesario incrementar α o aumentar el tamaño de la muestra.

Hasta ahora gran parte del análisis de la prueba de hipótesis se ha enfocado en los principios y las definiciones. En las secciones que siguen seremos más específicos y

clasificaremos las hipótesis en categorías. También estudiaremos pruebas de hipótesis sobre varios parámetros de interés. Comenzamos estableciendo la diferencia entre hipótesis unilaterales y bilaterales.

Pruebas de una y dos colas

Una prueba de cualquier hipótesis estadística donde la alternativa es **unilateral**, como

$$H_0: \theta = \theta_0,$$
$$H_1: \theta > \theta_0,$$

o quizás

$$H_0: \theta = \theta_0,$$
$$H_1: \theta < \theta_0,$$

se denomina **prueba de una sola cola**. Anteriormente en esta sección se hizo referencia al **estadístico de prueba** para una hipótesis. Por lo general la región crítica para la hipótesis alternativa $\theta > \theta_0$ yace en la cola derecha de la distribución del estadístico de prueba, en tanto que la región crítica para la hipótesis alternativa $\theta < \theta_0$ yace por completo en la cola izquierda. (En cierto sentido el símbolo de desigualdad señala la dirección en donde se encuentra la región crítica). En el experimento de la vacuna se utilizó una prueba de una sola cola para probar la hipótesis $p = 1/4$ contra la alternativa unilateral $p > 1/4$ para la distribución binomial. La región crítica de una sola cola por lo general es evidente; el lector debería visualizar el comportamiento del estadístico de prueba y observar la *señal* evidente que produciría evidencia que respalde la hipótesis alternativa.

La prueba de cualquier hipótesis alternativa donde la alternativa es **bilateral**, como

$$H_0: \theta = \theta_0,$$
$$H_1: \theta \neq \theta_0,$$

se denomina **prueba de dos colas**, ya que la región crítica se divide en dos partes, a menudo con probabilidades iguales en cada cola de la distribución del estadístico de prueba. La hipótesis alternativa $\theta \neq \theta_0$ establece que $\theta < \theta_0$ o que $\theta > \theta_0$. Se utilizó una prueba de dos colas para probar la hipótesis nula $\mu = 68$ kilogramos contra la alternativa bilateral $\mu \neq 68$ kilogramos en el ejemplo de la población continua de los pesos de estudiantes.

¿Cómo se eligen las hipótesis nula y alternativa?

Con frecuencia la hipótesis nula H_0 se plantea usando el *signo de igualdad.* Con este método se observa claramente cómo se controla la probabilidad de cometer un error tipo I. Sin embargo, hay situaciones en que "no rechazar H_0" implica que el parámetro θ podría ser cualquier valor definido por el complemento natural de la hipótesis alternativa. Por ejemplo, en el caso de la vacuna, donde la hipótesis alternativa es $H_1: p > 1/4$, es muy posible que el no rechazo de H_0 no pueda descartar un valor de p menor que 1/4. Sin embargo, es evidente que en el caso de las pruebas de una cola la consideración más importante es el planteamiento de la alternativa.

La decisión de plantear una prueba de una cola o una de dos colas depende de la conclusión que se obtenga si se rechaza H_0. La ubicación de la región crítica sólo se puede determinar después de que se plantea H_1. Por ejemplo, al probar una medicina nueva se establece la hipótesis de que no es mejor que las medicinas similares que actualmente hay en el mercado y se prueba contra la hipótesis alternativa de que la medicina nueva es mejor. Esta hipótesis alternativa dará como resultado una prueba de una sola cola, con la región crítica en la cola derecha. Sin embargo, si deseamos comparar una nueva técnica de enseñanza con el procedimiento convencional del salón de clases, la hipótesis alternativa debe permitir que el nuevo método sea inferior o superior al procedimiento convencional. Por lo tanto, la prueba sería de dos colas con la región crítica dividida en partes iguales, de manera que caiga en los extremos de las colas izquierda y derecha de la distribución de nuestro estadístico.

Ejemplo 10.1: Un fabricante de cierta marca de cereal de arroz afirma que el contenido promedio de grasa saturada no excede a 1.5 gramos por porción. Plantee las hipótesis nula y alternativa que se utilizarán para probar esta afirmación y establezca en dónde se localiza la región crítica.

***Solución*:** La afirmación del fabricante se rechazará sólo si μ es mayor que 1.5 miligramos y no se rechazará si μ es menor o igual que 1.5 miligramos. Entonces, probamos

$$H_0: \mu = 1.5,$$
$$H_1: \mu > 1.5.$$

El hecho de no rechazar H_0 no descarta valores menores que 1.5 miligramos. Como tenemos una prueba de una cola, el símbolo mayor indica que la región crítica reside por completo en la cola derecha de la distribución de nuestro estadístico de prueba $\bar{X}$. ■

Ejemplo 10.2: Un agente de bienes raíces afirma que 60% de todas las viviendas privadas que se construyen actualmente son casas con tres dormitorios. Para probar esta afirmación se inspecciona una muestra grande de viviendas nuevas. Se registra la proporción de las casas con 3 dormitorios y se utiliza como estadístico de prueba. Plantee las hipótesis nula y alternativa que se utilizarán en esta prueba y determine la ubicación de la región crítica.

***Solución*:** Si el estadístico de prueba fuera considerablemente mayor o menor que $p = 0.6$, rechazaríamos la afirmación del agente. En consecuencia, deberíamos plantear las siguientes hipótesis:

$$H_0: p = 0.6,$$
$$H_1: p \neq 0.6.$$

La hipótesis alternativa implica una prueba de dos colas con la región crítica dividida por igual en ambas colas de la distribución de $\hat{P}$, nuestro estadístico de prueba. ■

10.3 Uso de valores *P* para la toma de decisiones en la prueba de hipótesis

Al probar hipótesis en las que el estadístico de prueba es discreto, la región crítica se podría elegir de manera arbitraria y determinar su tamaño. Si α es demasiado grande, se reduce haciendo un ajuste en el valor crítico. Quizá sea necesario aumentar el tamaño

de la muestra para compensar la disminución que ocurre de manera automática en la potencia de la prueba.

Por generaciones enteras de análisis estadístico se ha vuelto costumbre elegir una α de 0.05 o 0.01 y seleccionar la región crítica de acuerdo con esto. Entonces, desde luego, el rechazo o no rechazo estrictos de H_0 dependerá de esa región crítica. Por ejemplo, si la prueba es de dos colas, α se fija a un nivel de significancia de 0.05 y el estadístico de prueba implica, digamos, la distribución normal estándar, entonces se observa un valor z de los datos y la región crítica es

$$z > 1.96 \quad \text{o} \quad z < -1.96,$$

donde el valor 1.96 corresponde a $z_{0.025}$ en la tabla A.3. Un valor de z en la región crítica sugiere la aseveración: "El valor del estadístico de prueba es significativo", el cual se puede traducir al lenguaje del caso. Por ejemplo, si la hipótesis es dada por

$$H_0: \mu = 10,$$
$$H_1: \mu \neq 10,$$

se puede decir: "La media difiere de manera significativa del valor 10".

Preselección de un nivel de significancia

Esta preselección de un nivel de significancia α tiene sus raíces en la filosofía de que se debe controlar el riesgo máximo de cometer un error tipo I. Sin embargo, este enfoque no explica los valores del estadístico de prueba que están "cercanos" a la región crítica. Suponga, por ejemplo, que en el caso de $H_0: \mu = 10$, contra $H_1: \mu \neq 10$, se observa un valor $z = 1.87$. En términos estrictos, con $\alpha = 0.05$ el valor no es significativo; pero el riesgo de cometer un error tipo I si se rechaza H_0 en este caso difícilmente se podría considerar grave. De hecho, en una situación de dos colas, el riesgo se cuantifica como

$$P = 2P(Z > 1.87 \text{ cuando } \mu = 10) = 2(0.0307) = 0.0614.$$

Como resultado, 0.0614 es la probabilidad de obtener un valor de z tan grande o mayor (en magnitud) que 1.87 cuando, de hecho, $\mu = 10$. Aunque esta evidencia en contra de H_0 no es tan firme como la que resultaría de un rechazo a un nivel $\alpha = 0.05$, se trata de información importante para el usuario. De hecho, el uso continuo de $\alpha = 0.05$ o 0.01 tan sólo es un resultado de lo que los estándares han transmitido por generaciones. **En la estadística aplicada los usuarios han adoptado de forma extensa el método del valor *P***. El método está diseñado para dar al usuario una alternativa (en términos de una probabilidad) a la mera conclusión de "rechazo" o "no rechazo". El cálculo del valor P también proporciona al usuario información importante cuando el valor z cae *dentro de la región crítica ordinaria.* Por ejemplo, si z es 2.73, resulta informativo para el usuario observar que

$$P = 2(0.0032) = 0.0064,$$

y, por consiguiente, el valor z es significativo a un nivel considerablemente menor que 0.05. Es importante saber que bajo la condición de H_0 un valor de $z = 2.73$ es un evento demasiado raro. A saber, un valor al menos tan grande en magnitud sólo ocurriría 64 veces en 10,000 experimentos.

Demostración gráfica de un valor *P*

Una manera muy simple de explicar gráficamente un valor *P* consiste en considerar dos muestras distintas. Suponga que se están considerando dos materiales para cubrir un tipo específico de metal con el fin de evitar la corrosión. Se obtienen especímenes y se cubre un grupo con el material 1 y otro grupo con el material 2. Los tamaños muestrales son $n_1 = n_2 = 10$ para cada muestra y la corrosión se mide en el porcentaje del área superficial afectada. La hipótesis plantea que las muestras provienen de distribuciones comunes con media $\mu = 10$. Supongamos que la varianza de la población es 1.0. Entonces, probamos

$$H_0 : \mu_1 = \mu_2 = 10.$$

Representemos con la figura 10.8 una gráfica de puntos de los datos. Los datos se colocan en la distribución determinada por la hipótesis nula. Supongamos que los datos "×" se refieren al material 1 y que los datos "∘" se refieren al material 2. Parece evidente que los datos realmente refutan la hipótesis nula. Pero, ¿cómo se podría resumir esto en un número? **El valor *P* se puede considerar simplemente como la probabilidad de obtener este conjunto de datos dado que las muestras provienen de la misma distribución**. Es evidente que esta probabilidad es muy pequeña, ¡digamos 0.00000001! Por consiguiente, el pequeño valor *P* evidentemente refuta H_0, y la conclusión es que las medias de la población son significativamente diferentes.

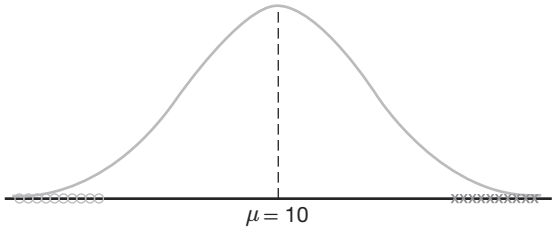


Figura 10.8: Datos que son probablemente generados de poblaciones que tienen dos medias diferentes.

El uso del método del valor *P* como auxiliar en la toma de decisiones es muy natural y casi todos los programas de cómputo que proporcionan el cálculo de pruebas de hipótesis ofrecen valores *P,* junto con valores del estadístico de prueba adecuado. La siguiente es una definición formal de un valor *P*.

Definición 10.5: Un **valor *P*** es el nivel (de significancia) más bajo en el que el valor observado del estadístico de prueba es significativo.

¿En qué difiere el uso de los valores *P* de la prueba de hipótesis clásica?

En este momento resulta tentador resumir los procedimientos que se asocian con la prueba de, digamos, H_0: $\theta = \theta_0$. Sin embargo, el estudiante que es novato en esta área deberá tener en cuenta que hay diferencias entre el enfoque y la filosofía del método

clásico de α fija, que tiene su momento más importante en la conclusión de "rechazar H_0" o "no rechazar H_0" y el método del valor P. En este último no se determina una α fija y las conclusiones se obtienen con base en el tamaño del valor P, según la apreciación subjetiva del ingeniero o del científico. Aun cuando los modernos programas de cómputo proporcionan valores P, es importante que el lector comprenda ambos enfoques para apreciar la totalidad de los conceptos. Por lo tanto, ofrecemos una breve lista con los pasos del procedimiento tanto para el método clásico como para el del valor P.

Aproximación a la prueba de hipótesis con probabilidad fija del error tipo I

1. Establezca las hipótesis nula y alternativa.
2. Elija un nivel de significancia α fijo.
3. Seleccione un estadístico de prueba adecuado y establezca la región crítica con base en α.
4. Rechace H_0 si el estadístico de prueba calculado está en la región crítica. De otra manera, no rechace H_0.
5. Saque conclusiones científicas y de ingeniería.

Prueba de significancia (método del valor P)

1. Establezca las hipótesis nula y alternativa.
2. Elija un estadístico de prueba adecuado.
3. Calcule el valor P con base en los valores calculados del estadístico de prueba.
4. Saque conclusiones con base en el valor P y los conocimientos del sistema científico.

En secciones posteriores de este capítulo y en los capítulos siguientes muchos ejemplos y ejercicios destacarán el método del valor P para obtener conclusiones científicas.

Ejercicios

10.1 Suponga que un alergólogo desea probar la hipótesis de que al menos 30% del público es alérgico a algunos productos de queso. Explique cómo el alergólogo podría cometer

a) un error tipo I;
b) un error tipo II.

10.2 Una socióloga se interesa en la eficacia de un curso de entrenamiento diseñado para lograr que más conductores utilicen los cinturones de seguridad en los automóviles.

a) ¿Qué hipótesis pone a prueba si comete un error tipo I al concluir de manera errónea que el curso de entrenamiento no es eficaz?
b) ¿Qué hipótesis pone a prueba si comete un error tipo II al concluir de forma errónea que el curso de entrenamiento es eficaz?

10.3 Se acusa a una empresa grande de discriminación en sus prácticas de contratación.

a) ¿Qué hipótesis se pone a prueba si un jurado comete un error tipo I al encontrar culpable a la empresa?
b) ¿Qué hipótesis se pone a prueba si un jurado comete un error tipo II al encontrar culpable a la empresa?

10.4 Un fabricante de telas considera que la proporción de pedidos de materia prima que llegan con retraso es $p = 0.6$. Si una muestra aleatoria de 10 pedidos indica que 3 o menos llegaron con retraso, la hipótesis de que $p = 0.6$ se debería rechazar a favor de la alternativa $p < 0.6$. Utilice la distribución binomial.

a) Calcule la probabilidad de cometer un error tipo I si la proporción verdadera es $p = 0.6$.
b) Calcule la probabilidad de cometer un error tipo II para las alternativas $p = 0.3$, $p = 0.4$ y $p = 0.5$.

10.5 Repita el ejercicio 10.4 pero suponga que se seleccionan 50 pedidos y que se define a la región crítica como $x \leq 24$, donde x es el número de pedidos en la muestra que llegaron con retraso. Utilice la aproximación normal.

10.6 Se estima que la proporción de adultos que vive en una pequeña ciudad que son graduados universitarios es $p = 0.6$. Para probar esta hipótesis se selecciona una muestra aleatoria de 15 adultos. Si el número de graduados en la muestra es cualquier número entre 6 y 12, no rechazaremos la hipótesis nula de que $p = 0.6$; de otro modo, concluiremos que $p \neq 0.6$.

a) Evalúe α suponiendo que $p = 0.6$. Utilice la distribución binomial.

b) Evalúe β para las alternativas $p = 0.5$ y $p = 0.7$.
c) ¿Es éste un buen procedimiento de prueba?

10.7 Repita el ejercicio 10.6 pero suponga que se seleccionan 200 adultos y que la región de no rechazo se define como $110 \leq x \leq 130$, donde x es el número de individuos graduados universitarios en la muestra. Utilice la aproximación normal.

10.8 En la publicación *Relief from Arthritis* de Thorsons Publishers, Ltd., John E. Croft afirma que más de 40% de los individuos que sufren de osteoartritis experimentan un alivio medible con un ingrediente producido por una especie particular de mejillón que se encuentra en la costa de Nueva Zelanda. Para probar esa afirmación se suministra el extracto de mejillón a un grupo de 7 pacientes con osteoartritis. Si 3 o más de los pacientes experimentan alivio, no rechazaremos la hipótesis nula de que $p = 0.4$; de otro modo, concluiremos que $p < 0.4$.
a) Evalúe α suponiendo que $p = 0.4$.
b) Evalúe β para la alternativa $p = 0.3$.

10.9 Una tintorería afirma que un nuevo removedor de manchas quitará más de 70% de las manchas en las que se aplique. Para verificar esta afirmación el removedor de manchas se utilizará sobre 12 manchas elegidas al azar. Si se eliminan menos de 11 de las manchas, no se rechazará la hipótesis nula de que $p = 0.7$; de otra manera, concluiremos que $p > 0.7$.
a) Evalúe α, suponiendo que $p = 0.7$.
b) Evalúe β para la alternativa $p = 0.9$.

10.10 Repita el ejercicio 10.9 pero suponga que se tratan 100 manchas y que la región crítica se define como $x > 82$, donde x es el número de manchas eliminadas.

10.11 Repita el ejercicio 10.8 pero suponga que el extracto de mejillón se administra a 70 pacientes y que la región crítica se define como $x < 24$, donde x es el número de pacientes con osteoartritis que experimentan alivio.

10.12 Se pregunta a una muestra aleatoria de 400 votantes en cierta ciudad si están a favor de un impuesto adicional de 4% sobre las ventas de gasolina con el fin de obtener los fondos que se necesitan con urgencia para la reparación de calles. Si más de 220 votantes, pero menos de 260 de ellos, favorecen el impuesto sobre las ventas, concluiremos que 60% de los votantes lo apoyan.
a) Calcule la probabilidad de cometer un error tipo I si 60% de los votantes están a favor del aumento de impuestos.
b) ¿Cuál es la probabilidad de cometer un error tipo II al utilizar este procedimiento de prueba si en realidad sólo 48% de los votantes está a favor del impuesto adicional a la gasolina?

10.13 Suponga que en el ejercicio 10.12 concluimos que 60% de los votantes está a favor del impuesto sobre las ventas de gasolina si más de 214 votantes, pero menos de 266 de ellos, lo favorecen. Demuestre que esta nueva región crítica tiene como resultado un valor más pequeño para α a costa de aumentar β.

10.14 Un fabricante desarrolla un nuevo sedal para pesca que, según afirma, tiene una resistencia media a la rotura de 15 kilogramos con una desviación estándar de 0.5 kilogramos. Para probar la hipótesis de que $\mu = 15$ kilogramos contra la alternativa de que $\mu < 15$ kilogramos se prueba una muestra aleatoria de 50 sedales. La región crítica se define como $\bar{x} < 14.9$.
a) Calcule la probabilidad de cometer un error tipo I cuando H_0 es verdadera.
b) Evalúe β para las alternativas $\mu = 14.8$ y $\mu = 14.9$ kilogramos.

10.15 En un restaurante de carnes una máquina de bebidas gaseosas se ajusta para que la cantidad de bebida que sirva se distribuya de forma aproximadamente normal, con una media de 200 mililitros y una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 9 bebidas y calculando el contenido promedio. Si $\bar{x}$ cae en el intervalo $191 < \bar{x} < 209$, se considera que la máquina opera de forma satisfactoria; de otro modo, se concluye que $\mu \neq 200$ mililitros.
a) Calcule la probabilidad de cometer un error tipo I cuando $\mu = 200$ mililitros.
b) Calcule la probabilidad de cometer un error tipo II cuando $\mu = 215$ mililitros.

10.16 Repita el ejercicio 10.15 para muestras de tamaño $n = 25$. Utilice la misma región crítica.

10.17 Se desarrolla un nuevo proceso de cura para cierto tipo de cemento que da como resultado una resistencia media a la compresión de 5000 kilogramos por centímetro cuadrado y una desviación estándar de 120 kilogramos. Para probar la hipótesis de que $\mu = 5000$ contra la alternativa de que $\mu < 5000$ se toma una muestra aleatoria de 50 piezas de cemento. La región crítica se define como $\bar{x} < 4970$.
a) Calcule la probabilidad de cometer un error tipo I cuando H_0 es verdadera.
b) Evalúe β para las alternativas $\mu = 4970$ y $\mu = 4960$.

10.18 Si graficamos las probabilidades de no rechazar H_0 que corresponden a diversas alternativas para μ (incluido el valor especificado para H_0) y conectamos todos los puntos mediante una curva suave, obtenemos la **curva característica de operación** del criterio de prueba o, simplemente, la curva CO. Observe que la probabilidad de no rechazar H_0 cuando es verdadera es simplemente $1 - \alpha$. Las curvas características de operación se utilizan con amplitud en aplicaciones industriales para proporcionar una muestra visual de los

méritos del criterio de prueba. Remítase al ejercicio 10.15 y calcule las probabilidades de no rechazar H_0 para los siguientes 9 valores de μ y grafique la curva CO: 184, 188, 192, 196, 200, 204, 208, 212 y 216.

10.4 Una sola muestra: pruebas respecto a una sola media

En esta sección consideramos de manera formal pruebas de hipótesis para una sola media de la población. Muchos de los ejemplos de las secciones anteriores incluyen pruebas sobre la media, por lo que el lector ya debería tener una idea de algunos de los detalles que aquí se describen.

Pruebas para una sola media (varianza conocida)

Primero deberíamos describir las suposiciones en las que se basa el experimento. El modelo para la situación subyacente se centra alrededor de un experimento con $X_1, X_2,..., X_n$, que representan una muestra aleatoria de una distribución con media μ y varianza $\sigma^2 > 0$. Considere primero la hipótesis

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0.$$

El estadístico de prueba adecuado se debe basar en la variable aleatoria $\bar{X}$. En el capítulo 8 se presentó el teorema del límite central, el cual establece en esencia que, sin importar la distribución de X, la variable aleatoria $\bar{X}$ tiene una distribución casi normal con media μ y varianza σ^2/n para muestras de tamaño razonablemente grande. Por consiguiente, $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$. Podemos determinar, entonces, una región crítica basada en el promedio muestral calculado $\bar{x}$. Ahora ya debería quedarle claro al lector que habrá una región crítica de dos colas para la prueba.

Estandarización de $\bar{X}$

Es conveniente estandarizar $\bar{X}$ e incluir de manera formal la variable aleatoria **normal estándar** Z, donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Sabemos que, *bajo* H_0, es decir, si $\mu = \mu_0$, entonces $\sqrt{n}\left(\bar{X} - \mu_0\right)/\sigma$ tiene una distribución $n(x; 0, 1)$ y, por lo tanto, la expresión

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

se puede utilizar para escribir una región de no rechazo adecuada. El lector debería tener en la mente que, formalmente, la región crítica se diseña para controlar α, la probabilidad de cometer un error tipo I. Debería ser evidente que se necesita una *señal de evidencia de dos colas* para apoyar H_1. Así, dado un valor calculado $\bar{x}$, la prueba formal implica rechazar H_0 si el *estadístico de prueba z* calculado cae en la región crítica que se describe a continuación.

Procedimiento de prueba para una sola media (varianza conocida)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{o} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

Si $-z_{\alpha/2} < z < z_{\alpha/2}$, no se rechaza H_0. El rechazo de H_0, desde luego, implica la aceptación de la hipótesis alternativa $\mu \neq \mu_0$. Con esta definición de la región crítica debería quedar claro que habrá α probabilidades de rechazar H_0 (al caer en la región crítica) cuando, en realidad, $\mu = \mu_0$.

Aunque es más fácil entender la región crítica escrita en términos de z, escribimos la misma región crítica en términos del promedio calculado $\bar{x}$. Lo siguiente se puede escribir como un procedimiento de decisión idéntico:

$$\textit{rechazar } H_0 \text{ si } \bar{x} < a \text{ o } \bar{x} > b,$$

donde

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

En consecuencia, para un nivel de significancia α, los valores críticos de la variable aleatoria z y $\bar{x}$ se presentan en la figura 10.9.

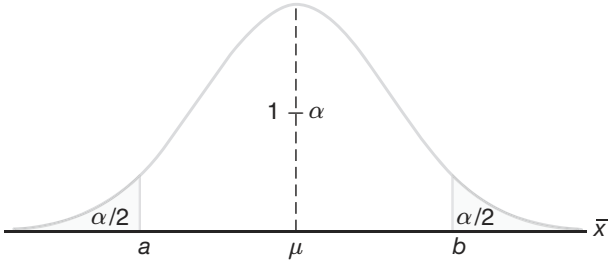


Figura 10.9: Región crítica para la hipótesis alternativa $\mu \neq \mu_0$.

Las pruebas de hipótesis unilaterales sobre la media incluyen el mismo estadístico que se describe en el caso bilateral. La diferencia, por supuesto, es que la región crítica sólo está en una cola de la distribución normal estándar. Por ejemplo, supongamos que buscamos probar

$$H_0: \mu = \mu_0,$$
$$H_1: \mu > \mu_0.$$

La señal que favorece H_1 proviene de *valores grandes* de z. Así, el rechazo de H_0 resulta cuando se calcula $z > z_\alpha$. Evidentemente, si la alternativa es $H_1: \mu < \mu_0$, la región crítica está por completo en la cola inferior, por lo que el rechazo resulta de $z < -z_\alpha$. Aunque en el caso de una prueba unilateral la hipótesis nula se puede escribir como $H_0: \mu \leq \mu_0$ o $H_0: \mu \geq \mu_0$, por lo general se escribe como $H_0: \mu = \mu_0$.

Los siguientes dos ejemplos ilustran pruebas de medias para el caso en el que se conoce σ.

Ejemplo 10.3: Una muestra aleatoria de 100 muertes registradas en Estados Unidos el año pasado reveló una vida promedio de 71.8 años. Si se supone una desviación estándar de la población de 8.9 años, ¿esto parece indicar que la vida media actual es mayor que 70 años? Utilice un nivel de significancia de 0.05.

***Solución*:**

1. H_0: $\mu = 70$ años.
2. H_1: $\mu > 70$ años.
3. $\alpha = 0.05$.
4. Región crítica: $z > 1.645$, donde $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
5. Cálculos: $\bar{x} = 71.8$ años, $\sigma = 8.9$ años, en consecuencia, $z = \dfrac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$.
6. Decisión: rechazar H_0 y concluir que la vida media actual es mayor que 70 años.

El valor P que corresponde a $z = 2.02$ es dado por el área de la región sombreada en la figura 10.10.

Si usamos la tabla A.3, tenemos

$$P = P(Z > 2.02) = 0.0217.$$

Como resultado, la evidencia a favor de H_1 es incluso más firme que la sugerida por un nivel de significancia de 0.05. ▟

Ejemplo 10.4: Un fabricante de equipo deportivo desarrolló un nuevo sedal para pesca sintético que, según afirma, tiene una resistencia media a la rotura de 8 kilogramos con una desviación estándar de 0.5 kilogramos. Pruebe la hipótesis de que $\mu = 8$ kilogramos contra la alternativa de que $\mu \neq 8$ kilogramos si se prueba una muestra aleatoria de 50 sedales y se encuentra que tienen una resistencia media a la rotura de 7.8 kilogramos. Utilice un nivel de significancia de 0.01.

***Solución*:**

1. H_0: $\mu = 8$ kilogramos.
2. H_1: $\mu \neq 8$ kilogramos.
3. $\alpha = 0.01$.
4. Región crítica: $z < -2.575$ y $z > 2.575$, donde $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
5. Cálculos: $\bar{x} = 7.8$ kilogramos, $n = 50$, en consecuencia, $z = \dfrac{7.8 - 8}{0.5/\sqrt{50}} = -2.83$.
6. Decisión: rechazar H_0 y concluir que la resistencia promedio a la rotura no es igual a 8 sino que, de hecho, es menor que 8 kilogramos.

Como la prueba en este ejemplo es de dos colas, el valor de P que se desea es el doble del área de la región sombreada en la figura 10.11 a la izquierda de $z = -2.83$. Por lo tanto, si usamos la tabla A.3, tenemos

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046,$$

que nos permite rechazar la hipótesis nula de que $\mu = 8$ kilogramos a un nivel de significancia menor que 0.01. ▟

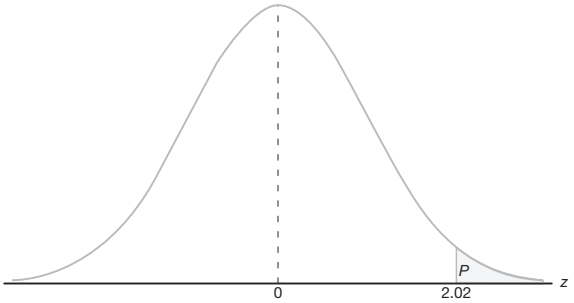


Figura 10.10: Valor *P* para el ejemplo 10.3.

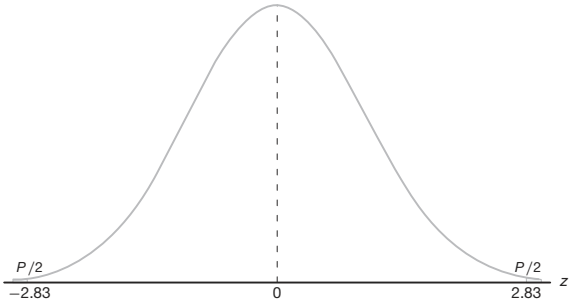


Figura 10.11: Valor *P* para el ejemplo 10.4.

Relación con la estimación del intervalo de confianza

El lector ya se habrá dado cuenta de que el método de la prueba de hipótesis para la inferencia estadística de este capítulo está muy relacionado con el método del intervalo de confianza del capítulo 9. La estimación del intervalo de confianza incluye el cálculo de límites dentro de los cuales es "razonable" que resida el parámetro en cuestión. Para el caso de una sola media de la población μ con σ^2 conocida, la estructura tanto de la prueba de hipótesis como de la estimación del intervalo de confianza se basa en la variable aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Resulta que la prueba de H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$ a un nivel de significancia α es *equivalente a calcular un intervalo de confianza del* $100(1-\alpha)\%$ *sobre* μ *y rechazar* H_0, *si* μ_0 *está fuera del intervalo de confianza.* Si μ_0 está dentro del intervalo de confianza, no se rechaza la hipótesis. La equivalencia es muy intuitiva y se puede ilustrar de manera muy simple. Recuerde que con un valor observado $\bar{x}$, no rechazar H_0 a un nivel de significancia α implica que

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

que es equivalente a

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

La equivalencia de la estimación del intervalo de confianza con la prueba de hipótesis se extiende a las diferencias entre dos medias, varianzas, cocientes de varianzas, etcétera. Como resultado, el estudiante de estadística no debería considerar la estimación del intervalo de confianza y la prueba de hipótesis como formas separadas de inferencia estadística. Considere el ejemplo 9.2 de la página 271. El intervalo de confianza del 95% sobre la media es dado por los límites (2.50, 2.70). Por consiguiente, con la misma información muestral, no se rechazará una hipótesis bilateral sobre μ que incluya cualquier valor hipotético entre 2.50 y 2.70. A medida que exploremos diferentes áreas de la prueba de hipótesis seguiremos aplicando la equivalencia a la estimación del intervalo de confianza.

Pruebas sobre una sola media (varianza desconocida)

Ciertamente sospecharíamos que las pruebas sobre una media de la población μ con σ^2 desconocida, como la estimación del intervalo de confianza, deberían incluir el uso de la distribución t de Student. En términos estrictos, la aplicación de la t de Student tanto para los intervalos de confianza como para la prueba de hipótesis se desarrolla bajo los siguientes supuestos. Las variables aleatorias $X_1, X_2, ..., X_n$ representan una muestra aleatoria de una distribución normal con μ y σ^2 desconocidas. Entonces, la variable aleatoria $\sqrt{n}(\bar{X} - \mu)/S$ tiene una distribución t de Student con $n - 1$ grados de libertad. La estructura de la prueba es idéntica a la del caso en el que se conoce σ, excepto que el valor σ en el estadístico de prueba se reemplaza con el estimado calculado de S y la distribución normal estándar se reemplaza con una distribución t.

El estadístico t para una prueba sobre una sola media (varianza desconocida)

Para la hipótesis bilateral

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0,$$

rechazamos H_0 a un nivel de significancia α cuando el estadístico t calculado

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

excede a $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$.

El lector debería recordar de los capítulos 8 y 9 que la distribución t es simétrica alrededor del valor cero. Así, esta región crítica de dos colas se aplica de manera similar a la del caso en que se conoce σ. Para la hipótesis bilateral a un nivel de significancia α se aplican las regiones críticas de dos colas. Para $H_1: \mu > \mu_0$ el rechazo resulta cuando $t > t_{\alpha,n-1}$. Para $H_1: \mu < \mu_0$ la región crítica es dada por $T < -t_{\alpha,n-1}$.

Ejemplo 10.5: El Edison Electric Institute publica cifras del número de kilowatts-hora que gastan anualmente varios aparatos electrodomésticos. Se afirma que una aspiradora gasta un promedio de 46 kilowatts-hora al año. Si una muestra aleatoria de 12 hogares, que se incluye en un estudio planeado, indica que las aspiradoras gastan un promedio de 42 kilowatts-hora al año con una desviación estándar de 11.9 kilowatts-hora, ¿esto sugiere que las aspiradoras gastan, en promedio, menos de 46 kilowatts-hora al año a un nivel de significancia de 0.05? Suponga que la población de kilowatts-hora es normal.

***Solución*:**

1. $H_0: \mu = 46$ kilowatts-hora.
2. $H_1: \mu < 46$ kilowatts-hora.
3. $\alpha = 0.05$.
4. Región crítica: $t < -1.796$, donde $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ con 11 grados de libertad.
5. Cálculos: $\bar{x} = 42$ kilowatts-hora, $s = 11.9$ kilowatts-hora y $n = 12$.
   En consecuencia,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decisión: no rechazar H_0 y concluir que el número promedio de kilowatts-hora que gastan al año las aspiradoras domésticas no es significativamente menor que 46. ■

Comentario sobre la prueba *t* de una sola muestra

Es probable que el lector haya observado que se mantiene la equivalencia de la prueba *t* de dos colas para una sola media y el cálculo de un intervalo de confianza sobre μ con σ reemplazada por *s*. Considere el ejemplo 9.5 de la página 275. En esencia, podemos ver ese cálculo como uno en el que encontramos todos los valores de μ_0, el volumen medio hipotético de contenedores de ácido sulfúrico para los que la hipótesis H_0: $\mu = \mu_0$ *no se rechazará* con $\alpha = 0.05$. De nuevo, esto es consistente con el planteamiento: "si nos basamos en la información muestral, son razonables los valores del volumen medio de la población entre 9.74 y 10.26 litros".

En este punto vale la pena destacar algunos comentarios respecto a la suposición de normalidad. Indicamos que cuando se conoce σ, el teorema del límite central permite utilizar un estadístico de prueba o un intervalo de confianza que se base en *Z*, la variable aleatoria normal estándar. En términos estrictos, por supuesto, el teorema del límite central y, por lo tanto, el uso de la distribución normal estándar, no se aplica a menos que se conozca σ. En el capítulo 8 se estudió el desarrollo de la distribución *t.* Ahí se estableció que la normalidad sobre $X_1, X_2, ..., X_n$ era una suposición subyacente. Entonces, *en sentido estricto,* no se deberían utilizar las tablas de *t* de Student de puntos porcentuales para pruebas o intervalos de confianza, a menos que se sepa que la muestra proviene de una población normal. En la práctica rara vez se puede suponer una σ conocida. Sin embargo, se dispondría de una buena estimación a partir de experimentos anteriores. Muchos libros de estadística sugieren que, cuando $n \geq 30$, es posible reemplazar con seguridad σ por *s* en el estadístico de prueba

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

con una población que tiene forma de campana y aun así utilizar las tablas *Z* para la región crítica adecuada. Aquí la implicación es que en realidad se recurre al teorema del límite central y que se confía en el hecho de que $s \approx \sigma$. Evidentemente, cuando se hace esto el resultado debe considerarse como una aproximación. Por consiguiente, un valor *P* calculado (de la distribución *Z*) de 0.15 puede ser 0.12 o quizá 0.17; o un intervalo de confianza calculado puede ser un intervalo de 93% de confianza en vez de un intervalo de 95% como se desea. Entonces, ¿qué sucede en las situaciones donde $n \leq 30$? El usuario no puede confiar en que *s* se acerque a σ, y para tomar en cuenta la inexactitud de la estimación el intervalo de confianza debería ser más ancho o el valor crítico de mayor magnitud. Los puntos porcentuales de la distribución *t* logran esto, pero sólo son correctos cuando la muestra proviene de una distribución normal. Desde luego, se pueden utilizar las gráficas de probabilidad normal para tener cierta idea de la desviación de la normalidad en un conjunto de datos.

Para muestras pequeñas a menudo resulta difícil detectar desviaciones de una distribución normal. (Las pruebas de la bondad del ajuste se presentan en una sección posterior de este capítulo). Para distribuciones en forma de campana de las variables aleatorias $X_1, X_2, ..., X_n$, es probable que el uso de la distribución *t* para pruebas o intervalos de confianza produzca resultados muy buenos. Cuando haya duda, el usuario debería recurrir a los procedimientos no paramétricos que se presentan en el capítulo 16.

Impresiones o salidas por computadora con comentarios para pruebas *t* de una sola muestra

Seguramente al lector le interesará ver comentarios impresos por computadora que muestren el resultado de una prueba *t* con una sola muestra. Suponga que un ingeniero se interesa en probar el sesgo en un medidor de pH. Se reúnen datos de una sustancia neutra (pH = 7.0). Se toma una muestra de las mediciones y los datos son los siguientes:

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

Entonces, es de interés probar

$$H_0: \mu = 7.0,$$

$$H_1: \mu \neq 7.0.$$

En este caso utilizamos el paquete de cómputo *MINITAB* para ilustrar el análisis del conjunto de datos anterior. Observe los componentes clave de la impresión o salida que se muestra en la figura 10.12. Desde luego, la media $\bar{y}$ es 7.0250, StDev es simplemente la desviación estándar de la muestra $s = 0.044$ y SE Mean es el error estándar estimado de la media, y se calcula como $s/\sqrt{n} = 0.0139$. El valor *t* es el cociente

$$(7.0250-7)/0.0139 = 1.80.$$

```
pH-meter
  7.07   7.00   7.10   6.97   7.00   7.03   7.01   7.01   6.98   7.08
MTB > Onet 'pH-meter'; SUBC>  Test7.

One-Sample T: pH-meter Test of mu = 7 vs not = 7
Variable   N   Mean    StDev  SE Mean        95% CI        T      P
pH-meter  10 7.02500 0.04403 0.01392 (6.99350,7.05650) 1.80  0.106
```

Figura 10.12: Impresión de *MINITAB* para la prueba *t* de una muestra para el medidor de pH.

El valor *P* de 0.106 sugiere resultados que no son concluyentes. No hay evidencia que sugiera un firme rechazo de H_0 (con base en una α de 0.05 o de 0.10), **ni se puede concluir con certeza que el medidor de pH esté libre de sesgo**. Observe que el tamaño de la muestra de 10 es muy pequeño. Un incremento en el tamaño de la muestra (quizás otro experimento) podría resolver las cosas. En la sección 10.6 aparece un análisis respecto al tamaño adecuado de la muestra.

10.5 Dos muestras: pruebas sobre dos medias

El lector deberá comprender la relación entre pruebas e intervalos de confianza y sólo puede confiar plenamente en los detalles que ofrece el material sobre el intervalo de confianza del capítulo 9. Las pruebas respecto a dos medias representan un conjunto de he-

rramientas analíticas muy importantes para el científico o el ingeniero. El procedimiento experimental es muy parecido al que se describe en la sección 9.8. Se extraen dos muestras aleatorias independientes de tamaños n_1 y n_2, respectivamente, de dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2. Sabemos que la variable aleatoria

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

tiene una distribución normal estándar. Suponemos aquí que n_1 y n_2 son suficientemente grandes, por lo que se aplica el teorema del límite central. Por supuesto, si las dos poblaciones son normales, el estadístico anterior tiene una distribución normal estándar incluso para n_1 y n_2 pequeñas. Evidentemente, si podemos suponer que $\sigma_1 = \sigma_2 = \sigma$, el estadístico anterior se reduce a

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

Los dos estadísticos anteriores sirven como base para el desarrollo de los procedimientos de prueba que incluyen dos medias. La equivalencia entre las pruebas y los intervalos de confianza, junto con los detalles técnicos implicados en las pruebas sobre una media, permiten que la transición a pruebas con dos medias sea sencilla.

La hipótesis bilateral sobre dos medias se escribe de manera muy general como

$$H_0: \mu_1 - \mu_2 = d_0.$$

Es evidente que la alternativa puede ser bilateral o unilateral. De nuevo, la distribución que se utiliza es la distribución del estadístico de prueba bajo H_0. Se calculan los valores $\bar{x}_1$ y $\bar{x}_2$, y para σ_1 y σ_2 conocidas, el estadístico de prueba es dado por

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

con una región crítica de dos colas en el caso de una alternativa bilateral. Es decir, se rechaza H_0 a favor de $H_1: \mu_1 - \mu_2 \neq d_0$, si $z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$. Las regiones críticas de una cola se utilizan en el caso de alternativas unilaterales. El lector debería estudiar, como antes, el estadístico de prueba y estar satisfecho de que para, digamos $H_1: \mu_1 - \mu_2 > d_0$, la señal que favorece H_1 provenga de valores grandes de z. Por consiguiente, se aplica la región crítica de la cola superior.

Varianzas desconocidas pero iguales

Las situaciones más comunes que implican pruebas sobre dos medias son aquellas con varianzas desconocidas. Si el científico interesado está dispuesto a suponer que ambas distribuciones son normales y que $\sigma_1 = \sigma_2 = \sigma$, se puede utilizar la *prueba t agrupada* (a menudo llamada prueba *t* de dos muestras). El estadístico de prueba (véase la sección 9.8) es dado por el siguiente procedimiento de prueba.

Prueba *t* agrupada de dos muestras

Para la hipótesis bilateral

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2,$$

rechazamos H_0 al nivel de significancia α cuando el estadístico t calculado

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

donde

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

excede a $t_{\alpha/2,\, n_1+n_2-2}$ o es menor que $-t_{\alpha/2,\, n_1+n_2-2}$.

Recuerde que en el capítulo 9 se explicó que los grados de libertad para la distribución t son un resultado del agrupamiento de la información de las dos muestras para estimar σ^2. Las alternativas unilaterales, como era de esperarse, sugieren regiones críticas unilaterales. Por ejemplo, para H_1: $\mu_1 - \mu_2 > d_0$, rechace H_1: $\mu_1 - \mu_2 = d_0$ cuando $t > t_{\alpha,\, n_1+n_2-2}$.

Ejemplo 10.6: Se llevó a cabo un experimento para comparar el desgaste por abrasivos de dos diferentes materiales laminados. Se probaron 12 piezas del material 1 exponiendo cada pieza a una máquina para medir el desgaste. Se probaron 10 piezas del material 2 de manera similar. En cada caso se observó la profundidad del desgaste. Las muestras del material 1 revelaron un desgaste promedio (codificado) de 85 unidades con una desviación estándar muestral de 4; en tanto que las muestras del material 2 revelaron un promedio de 81 y una desviación estándar muestral de 5. ¿Podríamos concluir, a un nivel de significancia de 0.05, que el desgaste abrasivo del material 1 excede al del material 2 en más de 2 unidades? Suponga que las poblaciones son aproximadamente normales con varianzas iguales.

***Solución*:** Representemos con μ_1 y μ_2 las medias de la población del desgaste abrasivo para el material 1 y el material 2, respectivamente.

1. $H_0: \mu_1 - \mu_2 = 2$.

2. $H_1: \mu_1 - \mu_2 > 2$.

3. $\alpha = 0.05$.

4. Región crítica: $t > 1.725$, donde $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ con $v = 20$ grados de libertad.

5. Cálculos:

$$\bar{x}_1 = 85, \quad s_1 = 4, \quad n_1 = 12,$$
$$\bar{x}_2 = 81, \quad s_2 = 5, \quad n_2 = 10.$$

En consecuencia,

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$
$$P = P(T > 1.04) \approx 0.16. \qquad \text{(Véase la tabla A.4).}$$

6. Decisión: no rechazar H_0. No podemos concluir que el desgaste abrasivo del material 1 excede al del material 2 en más de 2 unidades. ∎

Varianzas desconocidas pero diferentes

Hay situaciones donde al analista **no** le es posible suponer que $\sigma_1 = \sigma_2$. De la sección 9.8 recuerde que, si las poblaciones son normales, el estadístico

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

tiene una distribución t aproximada con grados de libertad aproximados

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Como resultado, el procedimiento de prueba consiste en *no rechazar* H_0 cuando

$$-t_{\alpha/2,v} < t' < t_{\alpha/2,v},$$

con v dado como antes. De nuevo, como en el caso de la prueba t agrupada, las alternativas unilaterales sugieren regiones críticas unilaterales.

Observaciones pareadas

Un estudio de la prueba t de dos muestras o el intervalo de confianza sobre la diferencia entre medias deberían sugerir la necesidad de un diseño experimental. Recuerde el análisis de las unidades experimentales en el capítulo 9, donde se sugirió que las condiciones de las dos poblaciones (a menudo denominadas como los dos tratamientos) se deberían asignar de manera aleatoria a las unidades experimentales. Esto se realiza para evitar resultados sesgados debido a las diferencias sistemáticas entre unidades experimentales. En otras palabras, en términos de la jerga para la prueba de hipótesis, es importante que la diferencia significativa que se encuentre entre las medias se deba a las diferentes condiciones de las poblaciones y no a las unidades experimentales en el estudio. Por ejemplo, considere el ejercicio 9.40 de la sección 9.9. Los 20 tallos desempeñan el papel de unidades experimentales. Diez de ellos se tratan con nitrógeno y 10 se dejan sin tratamiento. Es muy importante que esta asignación a los tratamientos "con nitrógeno" y "sin nitrógeno" sea aleatoria para garantizar que las diferencias sistemáticas entre los tallos no interfieran con una comparación válida entre las medias.

En el ejemplo 10.6 el momento de la medición es la opción más probable de la unidad experimental. Las 22 piezas de material se deberían medir en orden aleatorio.

Necesitamos protegernos contra la posibilidad de que las mediciones del desgaste que se realicen casi al mismo tiempo tiendan a dar resultados similares. *No se esperan* **diferencias sistemáticas** (no aleatorias) **en las unidades experimentales**. Sin embargo, las asignaciones aleatorias protegen contra el problema.

Las referencias a la planeación de experimentos, aleatorización, elección del tamaño de la muestra, etcétera, continuarán influyendo en gran parte del desarrollo en los capítulos 13, 14 y 15. Cualquier científico o ingeniero cuyo interés resida en el análisis de datos reales debería estudiar este material. La prueba *t* agrupada se amplía en el capítulo 13 para cubrir más de dos medias.

La prueba de dos medias se puede llevar a cabo cuando los datos están en forma de observaciones pareadas, como se estudió en el capítulo 9. En esta estructura de pareado las condiciones de las dos poblaciones (tratamientos) se asignan de forma aleatoria dentro de unidades homogéneas. El cálculo del intervalo de confianza para $\mu_1 - \mu_2$ en la situación con observaciones pareadas se basa en la variable aleatoria

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}},$$

donde $\bar{D}$ y S_d son variables aleatorias que representan la media muestral y la desviación estándar de las diferencias de las observaciones en las unidades experimentales. Como en el caso de la *prueba t agrupada,* la suposición es que las observaciones de cada población son normales. Este problema de dos muestras se reduce en esencia a un problema de una muestra utilizando las diferencias calculadas $d_1, d_2, \ldots, d_n$. Por consiguiente, la hipótesis se reduce a

$$H_0: \ \mu_D = d_0.$$

El estadístico de prueba calculado es dado entonces por

$$t = \frac{\overline{d} - d_0}{s_d/\sqrt{n}}.$$

Las regiones críticas se construyen usando la distribución *t* con $n - 1$ grados de libertad.

El problema de la interacción en una prueba *t* pareada

El siguiente estudio de caso no sólo ilustra el uso de la prueba *t* pareada, sino que el análisis revelará mucho sobre las dificultades que surgen cuando ocurre una interacción entre los tratamientos y las unidades experimentales en la estructura de la *t* pareada. Recuerde que la interacción entre factores se presentó en la sección 1.7, en un análisis de los tipos generales de estudios estadísticos. El concepto de interacción será un tema importante desde el capítulo 13 hasta el 15.

Existen ciertos tipos de pruebas estadísticas en los que la existencia de una interacción produce dificultades. Un ejemplo es la prueba *t* pareada. En la sección 9.9 se utilizó la estructura pareada en el cálculo de un intervalo de confianza sobre la diferencia entre dos medias, y se reveló la ventaja del pareado para situaciones en que las unidades experimentales son homogéneas. El pareado produce una reducción en σ_D, la desviación estándar de una diferencia $D_i = X_{1i} - X_{2i}$, como se explicó en la sección 9.9. Si hay una interacción entre los tratamientos y las unidades experimentales, la ventaja lograda

mediante el pareado se podría reducir de manera sustancial. Por consiguiente, en el ejemplo 9.13 de la página 293 la suposición de la ausencia de interacción permitió que la diferencia en los niveles medios de TCDD (plasma contra tejido adiposo) fuera la misma en todos los veteranos. Un vistazo rápido a los datos sugiere que no hay una violación significativa de los supuestos de ausencia de interacción.

Para demostrar cómo influye la interacción en Var(D) y, por lo tanto, en la calidad de la prueba t pareada, es aleccionador revisar la i-ésima diferencia dada por $D_i = X_{1i} - X_{2i} = (\mu_1 - \mu_2) + (\epsilon_1 - \epsilon_2)$, donde X_{1i} y X_{2i} se toman de la i-ésima unidad experimental. Si la unidad pareada es homogénea, los errores en X_{1i} y en X_{2i} serán similares y no independientes. En el capítulo 9 señalamos que la covarianza positiva entre los errores da como resultado una Var(D) reducida. Por consiguiente, el tamaño de la diferencia en los tratamientos y la relación entre los errores en X_{1i} y X_{2i}, a los que contribuye la unidad experimental, tenderán a permitir la detección de una diferencia significativa.

¿Qué condiciones resultan en una interacción?

Consideremos una situación en la que las unidades experimentales no son homogéneas. Más bien, considere la i-ésima unidad experimental con las variables aleatorias X_{1i} y X_{2i} que no son similares. Sean ϵ_{1i} y ϵ_{2i} variables aleatorias que representan los errores en los valores X_{1i} y X_{2i}, respectivamente, en la unidad i-ésima. Así, podemos escribir

$$X_{1i} = \mu_1 + \epsilon_{1i} \text{ y } X_{2i} = \mu_2 + \epsilon_{2i}.$$

Los errores con valor esperado cero podrían tender a provocar que los valores de respuesta X_{1i} y X_{2i} se muevan en direcciones opuestas, dando como resultado un valor negativo para Cov(ϵ_{1i}, ϵ_{2i}) y, por ende, un valor negativo para Cov(X_{1i}, X_{2i}). En realidad, el modelo se podría volver aún más complicado por el hecho de que $\sigma_1^2 = \text{Var}(\epsilon_{1i}) \neq \sigma_2^2 = \text{Var}(\epsilon_{2i})$. Los parámetros de la varianza y la covarianza podrían variar entre las n unidades experimentales. Así, a diferencia del caso con homogeneidad, D_i tenderá a ser muy diferente en todas las unidades experimentales debido a la naturaleza heterogénea de la diferencia en $\epsilon_1 - \epsilon_2$ entre las unidades. Esto produce la interacción entre los tratamientos y las unidades. Además, para una unidad experimental específica (véase el teorema 4.9),

$$\sigma_D^2 = \text{Var}(D) = \text{Var}(\epsilon_1) + \text{Var}(\epsilon_2) - 2\,\text{Cov}(\epsilon_1, \epsilon_2)$$

está inflado por el término negativo de covarianza, de manera que la ventaja lograda por el pareado en el caso de la unidad homogénea se pierde en el caso que aquí se describe. En tanto que la inflación en Var(D) variará de un caso a otro, en algunas situaciones existe el peligro de que el aumento en la varianza neutralice cualquier diferencia que exista entre μ_1 y μ_2. Desde luego, un valor grande de $\bar{d}$ en el estadístico t podría reflejar una diferencia en el tratamiento que compense el estimado inflado de la varianza s_d^2.

Estudio de caso 10.1: **Datos de muestra de sangre:** En un estudio realizado en el Departamento de Silvicultura y Fauna de Virginia Tech, J. A. Wesson examinó la influencia del fármaco *succinylcholine* sobre los niveles de circulación de andrógenos en la sangre. Se obtuvieron muestras de sangre de venados salvajes inmediatamente después de recibir una inyección intramuscular de *succinylcholine* con dardos de un rifle de caza. Treinta minutos después se obtuvo una segunda muestra de sangre y después los venados fueron liberados. Los

niveles de andrógenos de 15 venados al momento de la captura y 30 minutos más tarde, medidos en nanogramos por mililitro (ng/mL), se presentan en la tabla 10.2.

Suponga que las poblaciones de niveles de andrógenos al momento de la inyección y 30 minutos después se distribuyen normalmente, y pruebe, a un nivel de significancia de 0.05, si las concentraciones de andrógenos se alteraron después de 30 minutos.

Tabla 10.2: Datos para el estudio de caso 10.1

	Andrógenos (ng/mL)		
Venado	**Al momento de la inyección**	**30 minutos después de la inyección**	d_i
1	2.76	7.02	4.26
2	5.18	3.10	−2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	−2.45
11	11.78	11.10	−0.68
12	3.90	3.74	−0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

***Solución*:** Sean μ_1 y μ_2 la concentración promedio de andrógenos al momento de la inyección y 30 minutos después, respectivamente. Procedemos como sigue:

1. H_0: $\mu_1 = \mu_2$ o $\mu_D = \mu_1 - \mu_2 = 0$.

2. H_1: $\mu_1 \neq \mu_2$ o $\mu_D = \mu_1 - \mu_2 \neq 0$.

3. $\alpha = 0.05$.

4. Región crítica: $t < -2.145$ y $t > 2.145$, donde $t = \frac{\bar{d} - d_0}{s_D/\sqrt{n}}$ con $v = 14$ grados de libertad.

5. Cálculos: La media muestral y la desviación estándar para las d_i son

$$\bar{d} = 9.848 \text{ y } s_d = 18.474.$$

Por lo tanto,

$$t = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06.$$

6. Aunque el estadístico t no es significativo al nivel 0.05, de la tabla A.4,

$$P = P(|T| > 2.06) \approx 0.06.$$

Como resultado, existe cierta evidencia de que hay una diferencia en los niveles medios circulantes de andrógenos. ∎

La suposición de la ausencia de interacción implicaría que el efecto sobre los niveles de andrógenos de los venados es casi el mismo en los datos de ambos tratamientos, es decir, en el momento de la inyección de *succinylcholine* y 30 minutos después. Esto se puede expresar cambiando los papeles de los dos factores; por ejemplo, la diferencia en los tratamientos es casi igual en todas las unidades, es decir, los venados. Ciertamente hay algunas combinaciones venado/tratamiento para las que parece ser válida la suposición de ausencia de interacción, pero difícilmente existen evidencias firmes de que las unidades experimentales sean homogéneas. Sin embargo, la naturaleza de la interacción y el incremento resultante en $\text{Var}(\bar{D})$ parecen estar dominados por una diferencia sustancial en los tratamientos. Esto también es demostrado por el hecho de que 11 de los 15 venados mostraron señales positivas para las d_i calculadas y las d_i negativas (para los venados 2, 10, 11 y 12) son pequeñas en magnitud comparadas con las 12 positivas. Por consiguiente, al parecer el nivel medio de andrógenos es significativamente más alto 30 minutos después de la inyección que en el momento en que se aplica, y las conclusiones podrían ser más firmes de lo que sugiere $p = 0.06$.

Impresiones por computadora con comentarios para pruebas *t* pareadas

La figura 10.13 presenta una impresión por computadora del *SAS* para una prueba *t* pareada usando los datos del estudio de caso 10.1. Observe que el listado se parece al de una prueba *t* de una sola muestra y, por supuesto, esto es con exactitud lo que se realizó, ya que la prueba busca determinar si $\bar{d}$ es significativamente diferente de cero.

```
                Analysis Variable : Diff

 N           Mean      Std Error   t Value   Pr > |t|
-------------------------------------------------------
15      9.8480000     4.7698699      2.06     0.0580
-------------------------------------------------------
```

Figura 10.13: Impresión por computadora del *SAS* de la prueba *t* pareada para los datos del estudio de caso 10.1.

Resumen de los procedimientos de prueba

Mientras completamos el desarrollo formal de pruebas sobre medias de la población, ofrecemos la tabla 10.3, que resume el procedimiento de prueba para los casos de una sola media y de dos medias. Observe el procedimiento aproximado cuando las distribuciones son normales y las varianzas se desconocen pero no se suponen iguales. Este estadístico se estudió en el capítulo 9.

10.6 Elección del tamaño de la muestra para la prueba de medias

En la sección 10.2 demostramos cómo el analista puede explotar las relaciones entre el tamaño de la muestra, el nivel de significancia α y la potencia de la prueba para alcanzar cierto estándar de calidad. En la mayoría de las circunstancias prácticas el experimento debería planearse y, de ser posible, elegir el tamaño de la muestra antes del proceso de recolección de datos. Por lo general el tamaño de la muestra se determina de modo que

Tabla 10.3: Pruebas relacionadas con medias

H_0	Valor del estadístico de prueba	H_1	Región crítica
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ conocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ desconocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 y σ_2 conocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ pero desconocidas $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ y desconocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ o $t' > t_{\alpha/2}$
$\mu_D = d_0$ observaciones pareadas	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$

permita lograr una buena potencia para una α fija y una alternativa específica fija. Esta alternativa fija puede estar en la forma de $\mu - \mu_0$ en el caso de una hipótesis que incluya una sola media o $\mu_1 - \mu_2$ en el caso de un problema que implique dos medias. Los casos específicos serán ilustrativos.

Suponga que deseamos probar la hipótesis

$$H_0: \mu = \mu_0,$$
$$H_1 = \mu > \mu_0,$$

con un nivel de significancia α, cuando se conoce la varianza σ^2. Para una alternativa específica, digamos, $\mu = \mu_0 + \delta$, en la figura 10.14 se muestra que la potencia de nuestra prueba es

$$1 - \beta = P(\bar{X} > a \text{ cuando } \mu = \mu_0 + \delta).$$

Por lo tanto,

$$\begin{aligned}\beta &= P(\bar{X} < a \text{ cuando } \mu = \mu_0 + \delta)\\ &= P\left[\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \frac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \text{ cuando } \mu = \mu_0 + \delta\right].\end{aligned}$$

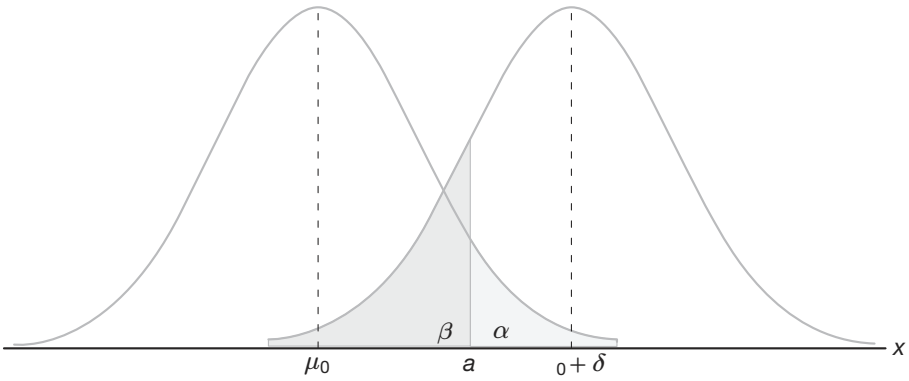


Figura 10.14: Prueba de $\mu = \mu_0$ contra $\mu = \mu_0 + \delta$.

Bajo la hipótesis alternativa $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}}$$

es la variable normal estándar Z. Por lo tanto,

$$\beta = P\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}\right) = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right),$$

de donde concluimos que

$$-z_\beta = z_\alpha - \frac{\delta\sqrt{n}}{\sigma},$$

y, en consecuencia,

$$\text{Elección del tamaño de la muestra: } n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2},$$

un resultado que también es verdadero cuando la hipótesis alternativa es $\mu < \mu_0$.

En el caso de una prueba de dos colas obtenemos la potencia $1 - \beta$ para una alternativa específica cuando

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Ejemplo 10.7: Suponga que deseamos probar la hipótesis

$$H_0: \mu = 68 \text{ kilogramos},$$
$$H_1: \mu > 68 \text{ kilogramos},$$

para los pesos de estudiantes hombres en cierta universidad usando un nivel de significancia $\alpha = 0.05$ cuando se sabe que $\sigma = 5$. Calcule el tamaño muestral que se requiere si la potencia de nuestra prueba debe ser 0.95 cuando la media real es 69 kilogramos.

***Solución*:** Como $\alpha = \beta = 0.05$, tenemos $z_\alpha = z_\beta = 1.645$. Para la alternativa $\beta = 69$ tomamos $\delta = 1$ y entonces,

$$n = \frac{(1.645 + 1.645)^2(25)}{1} = 270.6.$$

Por lo tanto, se requieren 271 observaciones si la prueba debe rechazar la hipótesis nula el 95% de las veces cuando, de hecho, μ es tan grande como 69 kilogramos. ∎

El caso de dos muestras

Se puede utilizar un procedimiento similar para determinar el tamaño de la muestra $n = n_1 = n_2$ que se requiere para una potencia específica de la prueba en que se comparan dos medias de la población. Por ejemplo, suponga que deseamos probar la hipótesis

$$H_0: \mu_1 - \mu_2 = d_0,$$
$$H_1: \mu_1 - \mu_2 \neq d_0,$$

cuando se conocen σ_1 y σ_2. Para una alternativa específica, digamos $\mu_1 - \mu_2 = d_0 + \delta$, en la figura 10.15 se muestra que la potencia de nuestra prueba es

$$1 - \beta = P\left(|\bar{X}_1 - \bar{X}_2| > a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta\right).$$

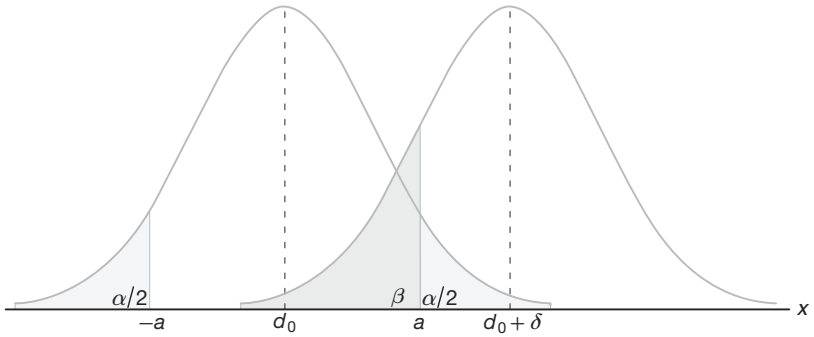


Figura 10.15: Prueba de $\mu_1 - \mu_2 = d_0$ contra $\mu_1 - \mu_2 = d_0 + \delta$.

Por lo tanto,

$$\begin{aligned}\beta &= P\left(-a < \bar{X}_1 - \bar{X}_2 < a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta\right)\\ &= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right.\\ &\qquad \left. < \frac{a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta\right].\end{aligned}$$

Con la hipótesis alternativa $\mu_1 - \mu_2 = d_0 + \delta$, el estadístico

$$\frac{\bar{X}_1 - \bar{X}_2 - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

es la variable normal estándar Z. Ahora bien, al escribir

$$-z_{\alpha/2} = \frac{-a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \quad \text{y} \quad z_{\alpha/2} = \frac{a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

tenemos

$$\beta = P\left[-z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < Z < z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right],$$

de donde concluimos que

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

y, por lo tanto,

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

Para la prueba de una sola cola, la expresión para el tamaño requerido de la muestra cuando $n = n_1 = n_2$ es

$$\text{Elección del tamaño de la muestra: } n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

Cuando se desconoce la varianza de la población (o varianzas en la situación de dos muestras), la elección del tamaño de la muestra no es directa. Al probar la hipótesis $\mu = \mu_0$ cuando el valor verdadero es $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0 + \delta)}{S/\sqrt{n}}$$

no sigue la distribución t, como se podría esperar, más bien sigue la **distribución *t* no central**. Sin embargo, existen tablas o gráficas que se basan en la distribución t no central para determinar el tamaño adecuado de la muestra, si se dispone de algún estimado de σ o si δ es un múltiplo de σ. La tabla A.8 proporciona los tamaños muestrales necesarios para controlar los valores de α y β para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu - \mu_0|}{\sigma}$$

en el caso de pruebas de una y de dos colas. En el caso de la prueba t de dos muestras en la que se desconocen las varianzas pero se suponen iguales, obtenemos los tamaños muestrales $n = n_1 = n_2$ necesarios para controlar los valores de α y β para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu_1 - \mu_2 - d_0|}{\sigma}$$

de la tabla A.9.

Ejemplo 10.8: Al comparar el comportamiento de dos catalizadores sobre el efecto del producto de una reacción se realiza una prueba t de dos muestras con $\alpha = 0.05$. Se considera que las

varianzas de los productos son iguales para los dos catalizadores. ¿De qué tamaño debe ser una muestra para cada catalizador si se desea probar la hipótesis

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2$$

si es esencial detectar una diferencia de 0.8σ entre los catalizadores con 0.9 de probabilidad?

***Solución*:** De la tabla A.9, con $\alpha = 0.05$ para una prueba de dos colas, $\beta = 0.1$ y

$$\Delta = \frac{|0.8\sigma|}{\sigma} = 0.8,$$

encontramos que el tamaño requerido de la muestra es $n = 34$.

En situaciones prácticas sería difícil forzar a un científico o a un ingeniero a hacer un compromiso sobre la información a partir de la cual se puede encontrar un valor de Δ. Se recuerda al lector que el valor Δ cuantifica el tipo de diferencia entre las medias que el científico considera importantes; es decir, una diferencia que se considere *significativa* desde un punto de vista científico, no estadístico. El ejemplo 10.8 ilustra cómo suele hacerse esta elección, a saber, mediante la selección de una fracción de σ. Evidentemente, si el tamaño de la muestra se basa en una elección de $|\delta|$, que es una fracción pequeña de σ, el tamaño muestral que resulta podría ser muy grande comparado con lo que permite el estudio.

10.7 Métodos gráficos para comparar medias

En el capítulo 1 se puso mucha atención a la presentación de datos en forma gráfica, como los diagramas de tallo y hojas y las gráficas de caja y bigote. En la sección 8.8 las gráficas de cuantiles y las gráficas normales cuantil-cuantil se utilizaron para brindar una "imagen" y resumir así un conjunto de datos experimentales. Muchos paquetes de cómputo producen representaciones gráficas. A medida que procedamos con otras formas de análisis de datos, por ejemplo, el análisis de regresión y el análisis de varianza, los métodos gráficos se vuelven aún más informativos.

Los auxiliares gráficos no se pueden utilizar como un reemplazo del propio procedimiento de prueba. En realidad, el valor del estadístico de prueba indica el tipo adecuado de evidencia en apoyo de H_0 o H_1. Sin embargo, una imagen ofrece una buena ilustración y a menudo es un mejor comunicador de evidencia para el beneficiario del análisis. Además, una imagen con frecuencia dejará claro por qué se encontró una diferencia significativa. La falla de una suposición importante se puede expresar mediante un resumen gráfico.

Para la comparación de medias, las gráficas de caja y bigote simultáneas proporcionan una imagen clara. El lector debería recordar que estas gráficas muestran el percentil 25, el percentil 75 y la mediana en un conjunto de datos. Además, las extensiones muestran los extremos en un conjunto de datos. Considere el ejercicio 10.40 al final de esta sección. Se midieron los niveles en plasma de ácido ascórbico en dos grupos de mujeres embarazadas: fumadoras y no fumadoras. En la figura 10.16 se observan las gráficas de caja y bigote para ambos grupos de mujeres y dos cosas son muy evidentes; al tomar en cuenta la variabilidad parece haber una diferencia despreciable en las medias muestrales. Además, parece que la variabilidad en los dos grupos es hasta cierto punto diferente. Desde luego, el analista debe tener en la mente las más bien considerables diferencias entre los tamaños muestrales en este caso.

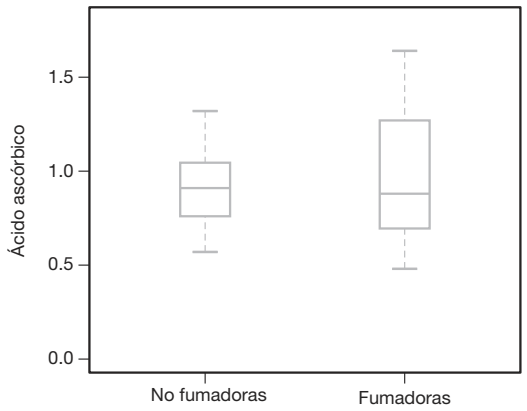


Figura 10.16: Dos gráficas de caja y bigote con los datos de ácido ascórbico para mujeres fumadoras y no fumadoras.

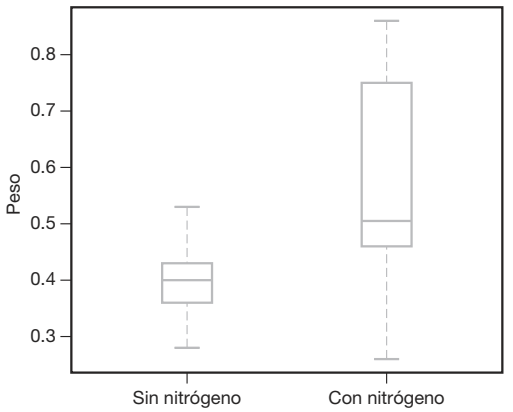


Figura 10.17: Dos gráficas de caja y bigote para los datos de los tallos.

Considere el ejercicio 9.40 de la sección 9.9. En la figura 10.17 se presenta la gráfica múltiple de caja y bigote para los datos de 10 tallos, de los cuales sólo la mitad recibió el tratamiento con nitrógeno. Tal gráfica revela una variabilidad menor para el grupo que no recibió nitrógeno. Además, la falta de traslape de las cajas sugiere una diferencia significativa entre los pesos medios de los tallos para los dos grupos. Parecería que la presencia de nitrógeno aumenta el peso de los tallos y quizás aumente la variabilidad en los pesos.

No existen reglas generales relacionadas con el momento cuando dos gráficas de caja y bigote brindan evidencia de diferencias significativas entre las medias. Sin embargo, una pauta aproximada es que si la línea del percentil 25 para una muestra excede a la línea de la mediana de la otra muestra, hay evidencia sólida de una diferencia entre las medias.

Se hará más énfasis en los métodos gráficos en un estudio de caso de la vida real que se presenta más adelante en este capítulo.

Impresiones por computadora con comentarios para pruebas *t* con dos muestras

Considere nuevamente el ejercicio 9.40 de la página 294, donde se reunieron datos de tallos que recibieron y no recibieron nitrógeno. Pruebe

$$H_0: \mu_{\text{NIT}} = \mu_{\text{NO}},$$
$$H_1: \mu_{\text{NIT}} > \mu_{\text{NO}},$$

donde las medias de la población indican los pesos medios. La figura 10.18 es una impresión por computadora con comentarios generados con el programa *SAS*. Observe que se presentan la desviación estándar y el error estándar muestrales para ambas muestras. También se incluye el estadístico t bajo la suposición de varianzas iguales y varianzas diferentes. En la gráfica de caja y bigote que se observa en la figura 10.17 en realidad parece que se transgrede la suposición de igualdad de varianzas. Un valor P de 0.0229 sugiere una conclusión de medias diferentes. Esto coincide con la información de diagnóstico que se presenta en la figura 10.18. A propósito, observe que t y t' son iguales en este caso, ya que $n_1 = n_2$.

```
                TTEST  Procedure
Variable  Weight
    Mineral     N       Mean  Std Dev  Std Err
No nitrogen    10     0.3990   0.0728   0.0230
   Nitrogen    10     0.5650   0.1867   0.0591

Variances      DF    t Value   Pr >|t|
    Equal      18       2.62    0.0174
  Unequal    11.7       2.62    0.0229

          Test the Equality of Variances
 Variable  Num DF     Den DF    F Value  Pr > F
   Weight       9          9       6.58  0.0098
```

Figura 10.18: Impresión del *SAS* para la prueba *t* de dos muestras.

Ejercicios

10.19 En un informe de investigación, Richard H. Weindruch, de la Escuela de Medicina de la UCLA, afirma que los ratones con una vida promedio de 32 meses vivirán hasta alrededor de 40 meses si 40% de las calorías en su dieta se reemplazan con vitaminas y proteínas. ¿Hay alguna razón para creer que $\mu < 40$, si 64 ratones que son sometidos a esa dieta tienen una vida promedio de 38 meses, con una desviación estándar de 5.8 meses? Utilice un valor *P* en su conclusión.

10.20 Una muestra aleatoria de 64 bolsas de palomitas con queso chedar pesan, en promedio, 5.23 onzas, con una desviación estándar de 0.24 onzas. Pruebe la hipótesis de que $\mu = 5.5$ onzas contra la hipótesis alternativa de que $\mu < 5.5$ onzas, al nivel de significancia de 0.05.

10.21 Una empresa de material eléctrico fabrica bombillas que tienen una duración que se distribuye de forma aproximadamente normal con una media de 800 horas y una desviación estándar de 40 horas. Pruebe la hipótesis de que $\mu = 800$ horas contra la alternativa de que $\mu \neq 800$ horas, si una muestra aleatoria de 30 bombillas tiene una duración promedio de 788 horas. Utilice un valor *P* en su respuesta.

10.22 En la revista *Hypertension* de la American Heart Association, investigadores reportan que los individuos que practican la meditación trascendental (MT) bajan su presión sanguínea de forma significativa. Si una muestra aleatoria de 225 hombres que practican la MT meditan 8.5 horas a la semana, con una desviación estándar de 2.25 horas, ¿esto sugiere que, en promedio, los hombres que utilizan la MT meditan más de 8 horas por semana? Cite un valor *P* en su conclusión.

10.23 Pruebe la hipótesis de que el contenido promedio de los envases de un lubricante específico es de 10 litros, si los contenidos de una muestra aleatoria de 10 envases son: 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 y 9.8 litros. Utilice un nivel de significancia de 0.01 y suponga que la distribución del contenido es normal.

10.24 La estatura promedio de mujeres en el grupo de primer año de cierta universidad ha sido, históricamente, de 162.5 centímetros, con una desviación estándar de 6.9 centímetros. ¿Existe alguna razón para creer que ha habido un cambio en la estatura promedio, si una muestra aleatoria de 50 mujeres del grupo actual de primer año tiene una estatura promedio de 165.2 centímetros? Utilice un valor *P* en su conclusión. Suponga que la desviación estándar permanece constante.

10.25 Se afirma que los automóviles recorren en promedio más de 20,000 kilómetros por año. Para probar tal afirmación se pide a una muestra de 100 propietarios de automóviles seleccionada de manera aleatoria que lleven un registro de los kilómetros que recorren. ¿Estaría usted de acuerdo con esta afirmación, si la muestra aleatoria indicara un promedio de 23,500 kilómetros y una desviación estándar de 3900 kilómetros? Utilice un valor *P* en su conclusión.

10.26 De acuerdo con un estudio sobre un régimen alimenticio, la ingesta elevada de sodio se relaciona con úlceras, cáncer estomacal y migrañas. El requerimiento humano de sal es de tan sólo 220 miligramos diarios, el cual se rebasa en la mayoría de las porciones individuales de cereales listos para comerse. Si una muestra aleatoria de 20 porciones similares de cierto cereal tiene un contenido medio de 244 miligramos de sodio y una desviación estándar de 24.5 miligramos, ¿esto sugiere, a un nivel de significancia de 0.05, que el contenido promedio de sodio para porciones individuales de ese cereal es mayor que 220 miligramos? Suponga que la distribución de contenidos de sodio es normal.

10.27 Un estudio de la Universidad de Colorado en Boulder revela que correr aumenta el porcentaje de la tasa metabólica basal (TMB) en mujeres ancianas. La TMB promedio de 30 ancianas corredoras fue 34.0%

más alta que la TMB promedio de 30 ancianas sedentarias, en tanto que las desviaciones estándar reportadas fueron de 10.5 y 10.2%, respectivamente. ¿Existe un aumento significativo en la TMB de las corredoras respecto a las sedentarias? Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales. Utilice un valor *P* en sus conclusiones.

10.28 De acuerdo con *Chemical Engineering,* una propiedad importante de la fibra es su absorbencia de agua. Se encontró que el porcentaje promedio de absorción de 25 pedazos de fibra de algodón seleccionados al azar es 20, con una desviación estándar de 1.5. Una muestra aleatoria de 25 pedazos de acetato reveló un porcentaje promedio de 12 con una desviación estándar de 1.25. ¿Existe evidencia sólida de que el porcentaje promedio de absorción de la población es significativamente mayor para la fibra de algodón que para el acetato? Suponga que el porcentaje de absorbencia se distribuye de forma casi normal y que las varianzas de la población en el porcentaje de absorbencia para las dos fibras son iguales. Utilice un nivel de significancia de 0.05.

10.29 La experiencia indica que el tiempo que requieren los estudiantes de último año de preparatoria para contestar una prueba estandarizada es una variable aleatoria normal con una media de 35 minutos. Si a una muestra aleatoria de 20 estudiantes de último año de preparatoria le toma un promedio de 33.1 minutos contestar esa prueba con una desviación estándar de 4.3 minutos, pruebe la hipótesis de que, a un nivel de significancia de 0.05, $\mu = 35$ minutos, contra la alternativa de que $\mu < 35$ minutos.

10.30 Una muestra aleatoria de tamaño $n_1 = 25$, tomada de una población normal con una desviación estándar $\sigma_1 = 5.2$, tiene una media $\bar{x}_1 = 81$. Una segunda muestra aleatoria de tamaño $n_2 = 36$, que se toma de una población normal diferente con una desviación estándar $\sigma_2 = 3.4$, tiene una media $\bar{x}_2 = 76$. Pruebe la hipótesis de que $\mu_1 = \mu_2$ contra la alternativa $\mu_1 \neq \mu_2$. Cite un valor *P* en su conclusión.

10.31 Un fabricante afirma que la resistencia promedio a la tensión del hilo *A* excede a la resistencia a la tensión promedio del hilo *B* en al menos 12 kilogramos. Para probar esta afirmación se pusieron a prueba 50 pedazos de cada tipo de hilo en condiciones similares. El hilo tipo *A* tuvo una resistencia promedio a la tensión de 86.7 kilogramos con una desviación estándar de 6.28 kilogramos; mientras que el hilo tipo *B* tuvo una resistencia promedio a la tensión de 77.8 kilogramos con una desviación estándar de 5.61 kilogramos. Pruebe la afirmación del fabricante usando un nivel de significancia de 0.05.

10.32 El *Amstat News* (diciembre de 2004) lista los sueldos medios de profesores asociados de estadística en instituciones de investigación, en escuelas de humanidades y en otras instituciones en Estados Unidos. Suponga que una muestra de 200 profesores asociados de instituciones de investigación tiene un sueldo promedio de $70,750 anuales con una desviación estándar de $6000. Suponga también que una muestra de 200 profesores asociados de otros tipos de instituciones tienen un sueldo promedio de $65,200 con una desviación estándar de $5000. Pruebe la hipótesis de que el sueldo medio de profesores asociados de instituciones de investigación es $2000 más alto que el de los profesores de otras instituciones. Utilice un nivel de significancia de 0.01.

10.33 Se llevó a cabo un estudio para saber si el aumento en la concentración de sustrato tiene un efecto apreciable sobre la velocidad de una reacción química. Con una concentración de sustrato de 1.5 moles por litro, la reacción se realizó 15 veces, con una velocidad promedio de 7.5 micromoles por 30 minutos y una desviación estándar de 1.5. Con una concentración de sustrato de 2.0 moles por litro, se realizaron 12 reacciones que produjeron una velocidad promedio de 8.8 micromoles por 30 minutos y una desviación estándar muestral de 1.2. ¿Hay alguna razón para creer que este incremento en la concentración de sustrato ocasiona un aumento en la velocidad media de la reacción de más de 0.5 micromoles por 30 minutos? Utilice un nivel de significancia de 0.01 y suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

10.34 Se realizó un estudio para determinar si los temas de un curso de física se comprenden mejor cuando éste incluye prácticas de laboratorio. Se seleccionaron estudiantes al azar para que participaran en un curso de tres semestres con una hora de clase sin prácticas de laboratorio o en un curso de cuatro semestres con una hora de clase con prácticas de laboratorio. En la sección con prácticas de laboratorio 11 estudiantes obtuvieron una calificación promedio de 85 con una desviación estándar de 4.7; mientras que en la sección sin prácticas de laboratorio 17 estudiantes obtuvieron una calificación promedio de 79 con una desviación estándar de 6.1. ¿Diría usted que el curso que incluyó prácticas de laboratorio aumentó la calificación promedio hasta en 8 puntos? Utilice un valor *P* en su conclusión y suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

10.35 Para indagar si un nuevo suero frena el desarrollo de la leucemia se seleccionan 9 ratones, todos en una etapa avanzada de la enfermedad. Cinco ratones reciben el tratamiento y cuatro no. Los tiempos de supervivencia, en años, a partir del momento en que comienza el experimento son los siguientes:

Con tratamiento	2.1	5.3	1.4	4.6	0.9
Sin tratamiento	1.9	0.5	2.8	3.1	

A un nivel de significancia de 0.05, ¿se puede decir que el suero es eficaz? Suponga que las dos poblaciones se distribuyen de forma normal con varianzas iguales.

10.36 Los ingenieros de una armadora de automóviles de gran tamaño están tratando de decidir si comprarán neumáticos de la marca *A* o de la marca *B* para sus modelos nuevos. Con el fin de ayudarlos a tomar una decisión se realiza un experimento en el que se usan 12 neumáticos de cada marca. Los neumáticos se utilizan hasta que se desgastan. Los resultados son los siguientes:

Marca *A*:

$$\bar{x}_1 = 37{,}900 \text{ kilómetros},$$
$$s_1 = 5100 \text{ kilómetros}.$$

Marca *B*:

$$\bar{x}_1 = 39{,}800 \text{ kilómetros},$$
$$s_2 = 5900 \text{ kilómetros}.$$

Pruebe la hipótesis de que no hay diferencia en el desgaste promedio de las 2 marcas de neumáticos. Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales. Use un valor *P*.

10.37 En el ejercicio 9.42 de la página 295 pruebe la hipótesis de que el ahorro de combustible de los camiones compactos Volkswagen, en promedio, excede al de los camiones compactos Toyota equipados de forma similar, que utilizan 4 kilómetros por litro. Utilice un nivel de significancia de 0.10.

10.38 Un investigador de la UCLA afirma que el promedio de vida de los ratones se puede prolongar hasta por 8 meses cuando se reducen las calorías en su dieta aproximadamente 40% desde el momento en que se destetan. Las dietas restringidas se enriquecen a niveles normales con vitaminas y proteínas. Suponga que a una muestra aleatoria de 10 ratones que tienen una vida promedio de 32.1 meses con una desviación estándar de 3.2 meses se les alimenta con una dieta normal, mientras que a una muestra aleatoria de 15 ratones que tienen un promedio de vida de 37.6 meses con una desviación estándar de 2.8 meses se les alimenta con la dieta restringida. A un nivel de significancia de 0.05 pruebe la hipótesis de que el promedio de vida de los ratones con esta dieta restringida aumenta 8 meses, contra la alternativa de que el aumento es menor de 8 meses. Suponga que las distribuciones de la esperanza de vida con las dietas regular y restringida son aproximadamente normales con varianzas iguales.

10.39 Los siguientes datos representan los tiempos de duración de películas producidas por 2 empresas cinematográficas:

Empresa	Tiempo (minutos)						
1	102	86	98	109	92		
2	81	165	97	134	92	87	114

Pruebe la hipótesis de que la duración promedio de las películas producidas por la empresa 2 excede al tiempo promedio de duración de las que produce la empresa 1 en 10 minutos, contra la alternativa unilateral de que la diferencia es de menos de 10 minutos. Utilice un nivel de significancia de 0.1 y suponga que las distribuciones de la duración son aproximadamente normales con varianzas iguales.

10.40 En un estudio realizado en Virginia Tech se compararon los niveles de ácido ascórbico en plasma en mujeres embarazadas fumadoras con los de mujeres no fumadoras. Para el estudio se seleccionaron 32 mujeres que estuvieran en los últimos 3 meses de embarazo, que no tuvieran padecimientos importantes y que sus edades fluctuaran entre los 15 y los 32 años. Antes de tomar muestras de 20 ml de sangre se pidió a las participantes que fueran en ayunas, que no tomaran sus suplementos vitamínicos y que evitaran alimentos con alto contenido de ácido ascórbico. A partir de las muestras de sangre se determinaron los siguientes valores de ácido ascórbico en el plasma de cada mujer, en miligramos por 100 mililitros:

Valores de ácido ascórbico en plasma

No fumadoras		Fumadoras
0.97	1.16	0.48
0.72	0.86	0.71
1.00	0.85	0.98
0.81	0.58	0.68
0.62	0.57	1.18
1.32	0.64	1.36
1.24	0.98	0.78
0.99	1.09	1.64
0.90	0.92	
0.74	0.78	
0.88	1.24	
0.94	1.18	

¿Existe suficiente evidencia para concluir que hay una diferencia entre los niveles de ácido ascórbico en plasma de mujeres fumadoras y no fumadoras? Suponga que los dos conjuntos de datos provienen de poblaciones normales con varianzas diferentes. Utilice un valor *P*.

10.41 El Departamento de Zoología de Virginia Tech llevó a cabo un estudio para determinar si existe una diferencia significativa en la densidad de organismos en dos estaciones diferentes ubicadas en Cedar Run, una corriente secundaria que se localiza en la cuenca del río Roanoke. El drenaje de una planta de tratamiento de aguas negras y el sobreflujo del estanque de sedimentación de la Federal Mogul Corporation entran al flujo cerca del nacimiento del río. Los siguientes datos proporcionan las medidas de densidad, en número de organismos por metro cuadrado, en las dos estaciones colectoras:

Número de organismos por metro cuadrado

Estación 1		Estación 2	
5030	4980	2800	2810
13,700	11,910	4670	1330
10,730	8130	6890	3320
11,400	26,850	7720	1230
860	17,660	7030	2130
2200	22,800	7330	2190
4250	1130		
15,040	1690		

A un nivel de significancia de 0.05, ¿podemos concluir que las densidades promedio en las dos estaciones son iguales? Suponga que las observaciones provienen de poblaciones normales con varianzas diferentes.

10.42 Cinco muestras de una sustancia ferrosa se usaron para determinar si existe una diferencia entre un análisis químico de laboratorio y un análisis de fluorescencia de rayos X del contenido de hierro. Cada muestra se dividió en dos submuestras y se aplicaron los dos tipos de análisis. A continuación se presentan los datos codificados que muestran los análisis de contenido de hierro:

	Muestra				
Análisis	**1**	**2**	**3**	**4**	**5**
Rayos X	2.0	2.0	2.3	2.1	2.4
Químico	2.2	1.9	2.5	2.3	2.4

Suponga que las poblaciones son normales y pruebe, al nivel de significancia de 0.05, si los dos métodos de análisis dan, en promedio, el mismo resultado.

10.43 De acuerdo con informes publicados, el ejercicio en condiciones de fatiga altera los mecanismos que determinan el desempeño. Se realizó un experimento con 15 estudiantes universitarios hombres, entrenados para realizar un movimiento horizontal continuo del brazo, de derecha a izquierda, desde un microinterruptor hasta una barrera, golpeando sobre la barrera en coincidencia con la llegada de una manecilla del reloj a la posición de las 6 en punto. Se registró el valor absoluto de la diferencia entre el tiempo, en milisegundos, que toma golpear sobre la barrera y el tiempo para que la manecilla alcance la posición de las 6 en punto (500 mseg). Cada participante ejecutó la tarea cinco veces en condiciones sin fatiga y con fatiga, y se registraron las siguientes sumas de las diferencias absolutas para las cinco ejecuciones:

	Diferencias absolutas de tiempo	
Sujeto	**Sin fatiga**	**Con fatiga**
1	158	91
2	92	59
3	65	215
4	98	226
5	33	223
6	89	91
7	148	92
8	58	177
9	142	134
10	117	116
11	74	153
12	66	219
13	109	143
14	57	164
15	85	100

Un aumento en la diferencia media absoluta de tiempo cuando la tarea se ejecuta en condiciones de fatiga apoyaría la afirmación de que el ejercicio, en condiciones de fatiga, altera el mecanismo que determina el desempeño. Suponga que las poblaciones se distribuyen normalmente y pruebe tal afirmación.

10.44 En un estudio realizado por el Departamento de Nutrición Humana y Alimentos del Virginia Tech se registraron los siguientes datos sobre los residuos de ácido sórbico en jamón, en partes por millón, inmediatamente después de sumergirlo en una solución de sorbato y después de 60 días de almacenamiento:

Residuos de ácido sórbico en jamón

Rebanada	Antes del almacenamiento	Después del almacenamiento
1	224	116
2	270	96
3	400	239
4	444	329
5	590	437
6	660	597
7	1400	689
8	680	576

Si se supone que las poblaciones se distribuyen normalmente, ¿hay suficiente evidencia, a un nivel de significancia de 0.05, para decir que la duración del almacenamiento influye en las concentraciones residuales de ácido sórbico?

10.45 El administrador de una empresa de taxis está tratando de decidir si el uso de neumáticos radiales en lugar de neumáticos regulares cinturados mejora el rendimiento de combustible. Se equipan 12 autos con neumáticos radiales y se conducen en un recorrido de prueba preestablecido. Sin cambiar a los conductores, los mismos autos se equipan con neumáticos regulares cinturados y se conducen nuevamente en el recorrido de prueba. Se registraron los siguientes datos sobre el consumo de gasolina, en kilómetros por litro:

Kilómetros por litro

Automóvil	Llantas radiales	Llantas cinturadas
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.7
11	6.1	6.0
12	5.2	4.9

¿Podemos concluir que los autos equipados con neumáticos radiales ahorran más combustible que aquellos equipados con neumáticos cinturados? Suponga que las poblaciones se distribuyen normalmente. Utilice un valor P en su conclusión.

10.46 En el ejercicio de repaso 9.91 de la página 313 utilice la distribución t para probar la hipótesis de que la dieta reduce el peso de un individuo en 4.5 kilogramos, en promedio, contra la hipótesis alternativa de que la diferencia media en peso es menor que 4.5 kilogramos. Utilice un valor P.

10.47 ¿Qué tan grande debería ser la muestra del ejercicio 10.20 para que la potencia de la prueba sea de 0.90, cuando la media verdadera es 5.20? Suponga que $\sigma = 0.24$.

10.48 Si la distribución del tiempo de vida en el ejercicio 10.19 es aproximadamente normal, ¿qué tan grande debería ser una muestra para que la probabilidad de cometer un error tipo II sea 0.1 cuando la media verdadera es 35.9 meses? Suponga que $\sigma = 5.8$ meses.

10.49 ¿Qué tan grande debería ser la muestra del ejercicio 10.24 para que la potencia de la prueba sea de 0.95 cuando la estatura promedio verdadera difiere de 162.5 en 3.1 centímetros? Utilice $\alpha = 0.02$.

10.50 ¿Qué tan grandes deberían ser las muestras del ejercicio 10.31 para que la potencia de la prueba sea de 0.95, cuando la diferencia verdadera entre los tipos de hilo A y B es 8 kilogramos?

10.51 ¿Qué tan grande debería ser la muestra del ejercicio 10.22 para que la potencia de la prueba sea de 0.8 cuando el tiempo promedio verdadero dedicado a la meditación excede al valor hipotético en 1.2 σ? Utilice $\alpha = 0.05$.

10.52 Se considera una prueba t a un nivel $\alpha = 0.05$ para probar

$$H_0: \mu = 14,$$
$$H_1: \mu \neq 14.$$

¿Qué tamaño de muestra se necesita para que la probabilidad de no rechazar de manera errónea H_0 sea 0.1 cuando la media de la población verdadera difiere de 14 en 0.5? A partir de una muestra preliminar estimamos que σ es 1.25.

10.53 En el Departamento de Medicina Veterinaria del Virginia Tech se llevó a cabo un estudio para determinar si la "resistencia" de una herida de incisión quirúrgica es afectada por la temperatura del bisturí. En el experimento se utilizaron 8 perros. Se hicieron incisiones "calientes" y "frías" en el abdomen de cada perro y se midió la resistencia. A continuación se presentan los datos resultantes.

Perro	Bisturí	Resistencia
1	Caliente	5120
1	Frío	8200
2	Caliente	10,000
2	Frío	8600
3	Caliente	10,000
3	Frío	9200
4	Caliente	10,000
4	Frío	6200
5	Caliente	10,000
5	Frío	10,000
6	Caliente	7900
6	Frío	5200
7	Caliente	510
7	Frío	885
8	Caliente	1020
8	Frío	460

a) Escriba una hipótesis adecuada para determinar si la resistencia de las incisiones realizadas con bisturí caliente difiere en forma significativa de la resistencia de las realizadas con bisturí frío.

b) Pruebe la hipótesis utilizando una prueba t pareada. Utilice un valor P en su conclusión.

10.54 Se utilizaron 9 sujetos en un experimento para determinar si la exposición a monóxido de carbono tiene un impacto sobre la capacidad respiratoria. Los datos fueron recolectados por el personal del Departamento de Salud y Educación Física del Virginia Tech y analizados en el Centro de Consulta Estadística en Hokie Land. Los sujetos fueron expuestos a cámaras de respiración, una de las cuales contenía una alta concentración de CO. Se realizaron varias mediciones de frecuencia respiratoria a cada sujeto en cada cámara. Los sujetos fueron expuestos a las cámaras de respiración en una secuencia aleatoria. Los siguientes datos representan la frecuencia respiratoria en número de respiraciones por minuto. Realice una prueba unilateral de la hipótesis de que la frecuencia respiratoria media es igual en los dos ambientes. Utilice $\alpha = 0.05$. Suponga que la frecuencia respiratoria es aproximadamente normal.

Sujeto	Con CO	Sin CO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

10.8 Una muestra: prueba sobre una sola proporción

Las pruebas de hipótesis que se relacionan con proporciones se requieren en muchas áreas. A los políticos les interesa conocer la fracción de votantes que los favorecerá en la siguiente elección. Todas las empresas manufactureras se preocupan por la proporción de artículos defectuosos cuando se realiza un embarque. Los jugadores dependen del conocimiento de la proporción de resultados que consideran favorables.

Consideraremos el problema de probar la hipótesis de que la proporción de éxitos en un experimento binomial es igual a algún valor específico. Es decir, probaremos la hipótesis nula H_0 de que $p = p_0$, donde p es el parámetro de la distribución binomial. La hipótesis alternativa puede ser una de las alternativas unilaterales o bilaterales usuales:

$$p < p_0, \quad p > p_0, \quad \text{o} \quad p \neq p_0.$$

La variable aleatoria adecuada sobre la que basamos nuestro criterio de decisión es la variable aleatoria binomial X; aunque también podríamos usar el estadístico $\hat{p} = X/n$. Los valores de X que están lejos de la media $\mu = np_0$ conducirán al rechazo de la hipótesis nula. Como X es una variable binomial discreta, es poco probable que se pueda establecer una región crítica cuyo tamaño sea *exactamente* igual a un valor preestablecido de α. Por esta razón es preferible, al trabajar con muestras pequeñas, basar nuestras decisiones en valores P. *P*ara probar la hipótesis

$$H_0: p = p_0,$$
$$H_1: p < p_0,$$

utilizamos la distribución binomial para calcular el valor P

$$P = P(X \leq x \text{ cuando } p = p_0).$$

El valor x es el número de éxitos en nuestra muestra de tamaño n. Si este valor P es menor o igual que α, nuestra prueba es significativa al nivel α y rechazamos H_0 a favor de H_1. De manera similar, para probar la hipótesis

$$H_0: p = p_0,$$
$$H_1: p > p_0,$$

al nivel de significancia α, calculamos

$$P = P(X \geq x \text{ cuando } p = p_0)$$

y rechazamos H_0 a favor de H_1 si este valor P es menor o igual que α. Finalmente, para probar la hipótesis

$$H_0: p = p_0,$$
$$H_1: p \neq p_0,$$

a un nivel de significancia α, calculamos

$$P = 2P(X \leq x \text{ cuando } p = p_0) \qquad \text{si } x < np_0$$

o

$$P = 2P(X \geq x \text{ cuando } p = p_0) \qquad \text{si } x > np_0$$

y rechazamos H_0 a favor de H_1 si el valor P calculado es menor o igual que α.

Los pasos para probar una hipótesis nula acerca de una proporción contra varias alternativas usando las probabilidades binomiales de la tabla A.1 son los siguientes:

Prueba de una proporción (muestras pequeñas)

1. H_0: $p = p_0$.
2. Una de las alternativas H_1: $p < p_0$, $p > p_0$ o $p \neq p_0$.
3. Elegir un nivel de significancia igual a α.
4. Estadístico de prueba: variable binomial X con $p = p_0$.
5. Cálculos: obtener x, el número de éxitos, y calcular el valor P adecuado.
6. Decisión: sacar las conclusiones apropiadas con base en el valor P.

Ejemplo 10.9: Un constructor afirma que en 70% de las viviendas que se construyen actualmente en la ciudad de Richmond, Virginia, se instalan bombas de calor. ¿Estaría de acuerdo con esta afirmación si una encuesta aleatoria de viviendas nuevas en esta ciudad revelara que 8 de 15 tienen instaladas bombas de calor? Utilice un nivel de significancia de 0.10.

***Solución*:**

1. H_0: $p = 0.7$.
2. H_1: $p \neq 0.7$.
3. $\alpha = 0.10$.
4. Estadístico de prueba: Variable binomial X con $p = 0.7$ y $n = 15$.
5. Cálculos: $x = 8$ y $np_0 = (15)(0.7) = 10.5$. Por lo tanto, de la tabla A.1, el valor P calculado es

$$P = 2P(X \leq 8 \text{ cuando } p = 0.7) = 2\sum_{x=0}^{8} b(x; 15, 0.7) = 0.2622 > 0.10.$$

6. Decisión: No rechazar H_0. Concluir que no hay razón suficiente para dudar de la afirmación del constructor. ∎

En la sección 5.2 aprendimos que cuando n es pequeña las probabilidades binomiales se pueden obtener de la fórmula binomial real o de la tabla A.1. Para n grande se requieren procedimientos de aproximación. Cuando el valor hipotético p_0 está muy cerca de 0 o de 1 se puede utilizar la distribución de Poisson con parámetro $\mu = np_0$. Sin embargo, para n grande por lo general se prefiere la aproximación de la curva normal, con los parámetros $\mu = np_0$ y $\sigma^2 = np_0q_0$, la cual es muy precisa, siempre y cuando p_0 no esté demasiado cerca de 0 o de 1. Si utilizamos la aproximación normal, el **valor z para probar $p = p_0$** es dado por

$$z = \frac{x - np_0}{\sqrt{np_0q_0}} = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}},$$

que es un valor de la variable normal estándar Z. Por consiguiente, para una prueba de dos colas al nivel de significancia α, la región crítica es $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$. Para la alternativa unilateral $p < p_0$, la región crítica es $z < -z_\alpha$, y para la alternativa $p > p_0$, la región crítica es $z > z_\alpha$.

Ejemplo 10.10: Se considera que un medicamento que se prescribe comúnmente para aliviar la tensión nerviosa tiene una eficacia de tan sólo 60%. Los resultados experimentales de un nuevo fármaco administrado a una muestra aleatoria de 100 adultos que padecían tensión nerviosa revelaron que 70 de ellos sintieron alivio. ¿Esta evidencia es suficiente para concluir que el nuevo medicamento es mejor que el que se prescribe comúnmente? Utilice un nivel de significancia de 0.05.

***Solución*:** **1.** H_0: $p = 0.6$.

2. H_1: $p > 0.6$.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$.

5. Cálculos: $x = 70$, $n = 100$, $\hat{p} = 70/100 = 0.7$ y

$$z = \frac{0.7-0.6}{\sqrt{(0.6)(0.4)/100}} = 2\ .04,\ P = P(Z > 2.04) < 0.0207.$$

6. Decisión: Rechazar H_0 y concluir que el nuevo fármaco es mejor.

10.9 Dos muestras: pruebas sobre dos proporciones

A menudo surgen situaciones en las que se desea probar la hipótesis de que dos proporciones son iguales. Por ejemplo, podemos tratar de mostrar evidencia de que la proporción de médicos que son pediatras en un estado es igual a la proporción de pediatras en otro estado. Quizás un individuo decida dejar de fumar sólo si se convence de que la proporción de fumadores con cáncer pulmonar excede a la proporción de no fumadores con ese tipo de cáncer.

En general, deseamos probar la hipótesis nula de que dos proporciones, o parámetros binomiales, son iguales. Es decir, probamos $p_1 = p_2$ contra una de las alternativas $p_1 < p_2$, $p_1 > p_2$, o $p_1 \neq p_2$. Desde luego, esto es equivalente a probar la hipótesis nula de que $p_1 - p_2 = 0$ contra una de las alternativas $p_1 - p_2 < 0$, $p_1 - p_2 > 0$ o $p_1 - p_2 \neq 0$. El estadístico sobre el que basamos nuestra decisión es la variable aleatoria $\hat{P}_1 - \hat{P}_2$. Se seleccionan al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones binomiales y se calcula la proporción de éxitos $\hat{P}_1$ y $\hat{P}_2$ para las dos muestras.

En la construcción de intervalos de confianza para p_1 y p_2 observamos, para n_1 y n_2 suficientemente grandes, que el estimador puntual $\hat{P}_1$ menos $\hat{P}_2$ estaba distribuido de forma casi normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Por lo tanto, es posible establecer la(s) región(es) crítica(s) usando la variable normal estándar

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}}.$$

Cuando H_0 es verdadera, podemos sustituir $p_1 = p_2 = p$ y $q_1 = q_2 = q$ (donde p y q son los valores comunes) en la fórmula anterior para Z y obtener la forma

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{pq(1/n_1 + 1/n_2)}}.$$

Sin embargo, para calcular un valor de Z debemos estimar los parámetros p y q que aparecen en el radical. Al agrupar los datos de ambas muestras el **estimado agrupado de la proporción** p es

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

donde x_1 y x_2 son el número de éxitos en cada una de las dos muestras. Al sustituir $\hat{p}$ por p y $\hat{q} = 1 - \hat{p}$ por q, el **valor z para probar $p_1 = p_2$** se determina a partir de la fórmula

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}.$$

Las regiones críticas para las hipótesis alternativas adecuadas se establecen como antes, utilizando puntos críticos de la curva normal estándar. En consecuencia, para la alternativa $p_1 \neq p_2$, al nivel de significancia α, la región crítica es $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$. Para una prueba donde la alternativa es $p_1 < p_2$, la región crítica será $z < -z_\alpha$; y cuando la alternativa es $p_1 > p_2$, la región crítica será $z > z_\alpha$.

Ejemplo 10.11: Se organizará una votación entre los residentes de una ciudad y el condado circundante para determinar si se aprueba una propuesta para la construcción de una planta química. Como el lugar en el que se propone construirla está dentro de los límites de la ciudad, muchos votantes del condado consideran que la propuesta será aprobada debido a la gran proporción de votantes que está a favor de que se construya. Se realiza una encuesta para determinar si hay una diferencia significativa en la proporción de votantes de la ciudad y los votantes del condado que favorecen la propuesta. Si 120 de 200 votantes de la ciudad favorecen la propuesta y 240 de 500 residentes del condado también lo hacen, ¿estaría usted de acuerdo en que la proporción de votantes de la ciudad que favorecen la propuesta es mayor que la proporción de votantes del condado? Utilice un nivel de significancia de $\alpha = 0.05$.

***Solución*:** Sean p_1 y p_2 las proporciones verdaderas de votantes en la ciudad y el condado, respectivamente, que favorecen la propuesta.

1. H_0: $p_1 = p_2$.
2. H_1: $p_1 > p_2$.
3. $\alpha = 0.05$
4. Región crítica: $z > 1.645$.
5. Cálculos:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = 0.60, \quad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = 0.48, \quad \text{y}$$
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 240}{200 + 500} = 0.51.$$

Por lo tanto,

$$z = \frac{0.60 - 0.48}{\sqrt{(0.51)(0.49)(1/200 + 1/500)}} = 2.9,$$
$$P = P(Z > 2.9) = 0.0019.$$

6. Decisión: Rechazar H_0 y estar de acuerdo en que la proporción de votantes de la ciudad a favor de la propuesta es mayor que la proporción de votantes del condado. ■

Ejercicios

10.55 Un experto en mercadotecnia de una empresa fabricante de pasta considera que 40% de los amantes de la pasta prefieren la lasagna. Si 9 de 20 amantes de la pasta eligen la lasagna sobre otras pastas, ¿qué se puede concluir acerca de la afirmación del experto? Utilice un nivel de significancia de 0.05.

10.56 Suponga que, en el pasado, 40% de todos los adultos estaban a favor de la pena capital. ¿Existe alguna razón para creer que la proporción de adultos que está a favor de la pena capital ha aumentado si, en una muestra aleatoria de 15 adultos, 8 están a favor de la pena capital? Utilice un nivel de significancia de 0.05.

10.57 Se está considerando utilizar un nuevo aparato de radar para cierto sistema de misiles de defensa. El sistema se verifica experimentando con una aeronave en la que se simula una situación en la que alguien muere y otra en la que no ocurre ninguna muerte. Si en 300 ensayos ocurren 250 muertes, al nivel de significancia de 0.04, acepte o rechace la afirmación de que la probabilidad de una muerte con el nuevo sistema no excede a la probabilidad de 0.8 del sistema que se utiliza actualmente.

10.58 Se cree que al menos 60% de los residentes de cierta área están a favor de una demanda de anexión de una ciudad vecina. ¿Qué conclusión extraería si sólo 110 en una muestra de 200 votantes están a favor de la demanda? Utilice un nivel de significancia de 0.05.

10.59 Una empresa petrolera afirma que en una quinta parte de las viviendas de cierta ciudad la gente utiliza petróleo como combustible para calentarlas. ¿Existen razones para creer que en menos de una quinta parte de las viviendas la gente utiliza este combustible para calentarlas si, en una muestra aleatoria de 1000 viviendas de esa ciudad, se encuentra que 136 utilizan petróleo como combustible? Utilice un valor P en su conclusión.

10.60 En cierta universidad se estima que a lo sumo 25% de los estudiantes van en bicicleta a la escuela. ¿Parece que ésta es una estimación válida si, en una muestra aleatoria de 90 estudiantes universitarios, se encuentra que 28 van en bicicleta a la escuela? Utilice un nivel de significancia de 0.05.

10.61 En un invierno con epidemia de influenza los investigadores de una conocida empresa farmacéutica encuestaron a los padres de 2000 bebés para determinar si el nuevo medicamento de la empresa era eficaz después de dos días. De 120 bebés que tenían influenza y que recibieron el medicamento, 29 se curaron en dos días o menos. De 280 bebés que tenían influenza pero no recibieron el fármaco, 56 se curaron en dos días o menos. ¿Hay alguna indicación significativa que apoye la afirmación de la empresa sobre la eficacia del medicamento?

10.62 En un experimento de laboratorio controlado, científicos de la Universidad de Minnesota descubrieron que 25% de cierta cepa de ratas sujetas a una dieta con 20% de grano de café y luego forzadas a consumir un poderoso químico causante de cáncer desarrollaron tumores cancerosos. Si el experimento se repite, y 16 de 48 ratas desarrollan tumores, ¿existen razones para creer que la proporción de ratas que desarrollan tumores cuando se someten a esta dieta se incrementa? Utilice un nivel de significancia de 0.05.

10.63 En un estudio que se realizó para estimar la proporción de residentes de cierta ciudad y sus suburbios que están a favor de que se construya una planta

de energía nuclear se encontró que 63 de 100 residentes urbanos están a favor de la construcción, mientras que sólo 59 de 125 residentes suburbanos la apoyan. ¿Hay una diferencia significativa entre la proporción de residentes urbanos y suburbanos que están a favor de que se construya la planta nuclear? Utilice un valor *P*.

10.64 En un estudio sobre la fertilidad de mujeres casadas, realizado por Martin O'Connell y Carolyn C. Rogers para la Oficina del Censo en 1979, se seleccionaron al azar dos grupos de mujeres casadas de entre 25 y 29 años de edad y sin hijos, y a cada una se le preguntó si planeaba tener un hijo en algún momento. Se seleccionó un grupo de mujeres con menos de dos años de casadas y otro de mujeres con cinco años de casadas. Suponga que 240 de 300 mujeres con menos de dos años de casadas planean tener un hijo algún día, en comparación con 288 de las 400 mujeres con cinco años de casadas. ¿Podemos concluir que la proporción de mujeres con menos de dos años de casadas que planean tener hijos es significativamente mayor que la proporción de mujeres con cinco años de casadas que también planean tenerlos? Utilice un valor *P*.

10.65 Una comunidad urbana quiere demostrar que la incidencia de cáncer de mama es mayor en su localidad que en una área rural vecina. (Se encontró que los niveles de PCB son más altos en el suelo de la comunidad urbana). Si descubre que en la comunidad urbana 20 de 200 mujeres adultas tienen cáncer de mama y que en la comunidad rural 10 de 150 mujeres adultas lo tienen, ¿podría concluir, con un nivel de significancia de 0.05, que el cáncer de mama prevalece más en la comunidad urbana?

10.66 **Proyecto de grupo:** Para este proyecto el grupo se debe dividir en parejas. Suponga que se supone que al menos 25% de los estudiantes de su universidad hacen más de dos horas de ejercicio por semana. Reúna datos de una muestra aleatoria de 50 estudiantes y pregunte a cada uno si se ejercita durante al menos dos horas por semana; luego haga los cálculos necesarios para rechazar o no rechazar la suposición anterior. Demuestre todo el procedimiento y utilice un valor *P* en sus conclusiones.

10.10 Pruebas de una y dos muestras referentes a varianzas

En esta sección estudiaremos la prueba de hipótesis relacionada con varianzas o desviaciones estándar de la población. No son poco comunes las aplicaciones de pruebas de una y dos muestras sobre varianzas. Los ingenieros y los científicos constantemente se enfrentan a estudios donde se les pide demostrar que las mediciones que tienen que ver con productos o procesos cumplen con las especificaciones que fijan los consumidores. Las especificaciones a menudo se cumplen si la varianza del proceso es suficientemente pequeña. También existe interés por experimentos que comparan métodos o procesos donde la reproducibilidad o variabilidad inherentes se deben comparar de manera formal. Además, para determinar si no se cumple la suposición de varianzas iguales, con frecuencia se aplica una prueba que compara dos varianzas antes de llevar a cabo una prueba *t* sobre dos medias.

Empecemos por considerar el problema de probar la hipótesis nula H_0 de que la varianza de la población σ^2 es igual a un valor específico σ_0^2 contra una de las alternativas comunes $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$ o $\sigma^2 \neq \sigma_0^2$. El estadístico apropiado sobre el que basamos nuestra decisión es el estadístico chi cuadrada del teorema 8.4, el cual se utilizó en el capítulo 9 para construir un intervalo de confianza para σ^2. Por lo tanto, si suponemos que la distribución de la población que se muestrea es normal, el valor de chi cuadrada para probar $\sigma^2 = \sigma_0^2$ es dado por

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

donde n es el tamaño de la muestra, s^2 es la varianza muestral y σ_0^2 es el valor de σ^2 dado por la hipótesis nula. Si H_0 es verdadera, χ^2 es un valor de la distribución chi cuadrada con $v = n - 1$ grados de libertad. En consecuencia, para una prueba de dos colas a un

nivel de significancia α, la región crítica es $\chi^2 < \chi^2_{1-\alpha/2}$ o $\chi^2 > \chi^2_{\alpha/2}$. Para la alternativa unilateral $\sigma^2 < \sigma_0^2$, la región crítica es $\chi^2 < \chi^2_{1-\alpha}$; y para la alternativa unilateral $\sigma^2 > \sigma_0^2$, la región crítica es $\chi^2 > \chi^2_{\alpha}$.

Robustez de la prueba χ^2 para la suposición de normalidad

Tal vez el lector se habrá dado cuenta de que varias pruebas dependen, al menos en teoría, de la suposición de normalidad. En general muchos procedimientos en estadística aplicada tienen fundamentos teóricos que dependen de la distribución normal. Estos procedimientos varían en el grado en que dependen de la suposición de la normalidad. A un procedimiento que es razonablemente insensible a esta suposición se le denomina **procedimiento robusto**, es decir, robusto para la normalidad. La prueba χ^2 sobre una sola varianza no es robusta en absoluto para la normalidad, es decir, el éxito práctico del procedimiento depende de la normalidad. Como resultado, el valor *P* calculado podría ser notoriamente diferente del valor *P* verdadero si la población de la que se toma la muestra no es normal. De hecho, resulta muy plausible que un valor *P* estadísticamente significativo no sea una verdadera señal de H_1: $\sigma \neq \sigma_0$, sino que un valor significativo sea el resultado de haber violado las suposiciones de normalidad. Por lo tanto, el analista debería utilizar esta prueba χ^2 específica con precaución.

Ejemplo 10.12: Un fabricante de baterías para automóvil afirma que la duración de sus baterías se distribuye de forma aproximadamente normal con una desviación estándar igual a 0.9 años. Si una muestra aleatoria de 10 de tales baterías tiene una desviación estándar de 1.2 años, ¿considera que $\sigma > 0.9$ años? Utilice un nivel de significancia de 0.05.

***Solución*:**

1. H_0: $\sigma^2 = 0.81$.

2. H_1: $\sigma^2 > 0.81$.

3. $\alpha = 0.05$.

4. Región crítica: En la figura 10.19 vemos que se rechaza la hipótesis nula cuando $\chi^2 > 16.919$, donde $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ con $v = 9$ grados de libertad.

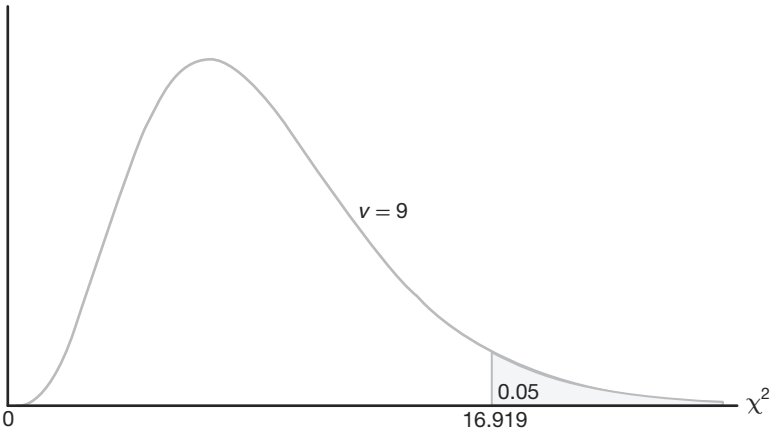


Figura 10.19: Región crítica para la hipótesis alternativa $\sigma > 0.9$.

5. Cálculos: $s^2 = 1.44$, $n = 10$ y

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.0, \qquad P \approx 0.07.$$

6. Decisión: El estadístico χ^2 no es significativo al nivel 0.05. Sin embargo, con base en el valor P de 0.07, hay evidencia de que $\sigma > 0.9$. ■

Consideremos ahora el problema de probar la igualdad de las varianzas σ_1^2 y σ_2^2 de dos poblaciones. Esto es, probaremos la hipótesis nula H_0 de que $\sigma_1^2 = \sigma_2^2$ contra una de las alternativas usuales

$$\sigma_1^2 < \sigma_2^2, \quad \sigma_1^2 > \sigma_2^2, \quad \text{o} \quad \sigma_1^2 \neq \sigma_2^2.$$

Para muestras aleatorias independientes de tamaños n_1 y n_2, respectivamente, de las dos poblaciones, **el valor *f* para probar $\sigma_1 = \sigma_2$** es el cociente

$$f = \frac{s_1^2}{s_2^2},$$

donde s_1^2 y s_2^2 son las varianzas calculadas de las dos muestras. Si las dos poblaciones se distribuyen de forma aproximadamente normal y la hipótesis nula es verdadera, de acuerdo con el teorema 8.8 el cociente $f = s_1^2 / s_2^2$ es un valor de la distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, las regiones críticas de tamaño α que corresponden a las alternativas unilaterales $\sigma_1^2 < \sigma_2^2$ y $\sigma_1^2 > \sigma_2^2$ son, respectivamente, $f < f_{1-\alpha}(v_1, v_2)$ y $f > f_{\alpha}(v_1, v_2)$. Para la alternativa bilateral $\sigma_1^2 \neq \sigma_2^2$ la región crítica es $f < f_{1-\alpha/2}(v_1, v_2)$ o $f > f_{\alpha/2}(v_1, v_2)$.

Ejemplo 10.13: Al probar la diferencia en el desgaste abrasivo de los dos materiales del ejemplo 10.6 supusimos que las dos varianzas de la población desconocidas eran iguales. ¿Se justifica tal suposición? Utilice un nivel de significancia de 0.10.

***Solución*:** Sean σ_1^2 y σ_2^2 las varianzas de la población para el desgaste abrasivo del material 1 y del material 2, respectivamente.

1. H_0: $\sigma_1^2 = \sigma_2^2$

2. H_1: $\sigma_1^2 \neq \sigma_2^2$

3. $\alpha = 0.10$.

4. Región crítica: En la figura 10.20 observamos que $f_{0.05}(11, 9) = 3.11$, y, usando el teorema 8.7, encontramos

$$f_{0.95}(11, 9) = \frac{1}{f_{0.05}(9, 11)} = 0.34.$$

Por lo tanto, se rechaza la hipótesis nula cuando $f < 0.34$ o $f > 3.11$, donde $f = s_1^2 / s_2^2$ con $\nu_1 = 11$ y $\nu_2 = 9$ grados de libertad.

5. Cálculos: $s_1^2 = 16$, $s_2^2 = 25$, por ende, $f = \frac{16}{25} = 0.64$.

6. Decisión: no rechazar H_0. Concluir que no hay suficiente evidencia de que las varianzas sean diferentes. ■

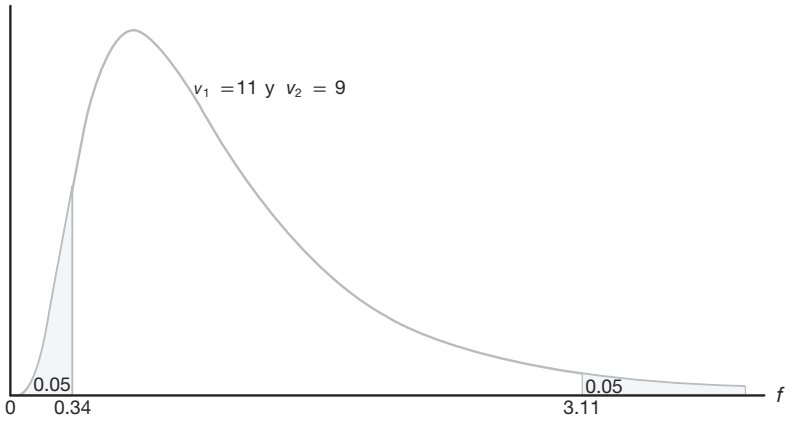


Figura 10.20: Región crítica para la hipótesis alternativa $\sigma_1^2 \neq \sigma_2^2$.

Prueba *F* para la prueba de varianzas con el SAS

La figura 10.18 de la página 356 presenta la impresión de una prueba *t* de dos muestras donde se comparan dos medias de los datos de los tallos en el ejercicio 9.40. La gráfica de caja y bigote que se observa en la figura 10.17 de la página 355 sugiere que las varianzas no son homogéneas y, por consiguiente, el estadístico t' y su valor *P* correspondiente son relevantes. Observe también que la impresión muestra el estadístico *F* para H_0: $\sigma_1 = \sigma_2$ con un valor *P* de 0.0098, que es evidencia adicional de que se debe esperar más variabilidad cuando se aplica el tratamiento con nitrógeno que cuando no se aplica.

Ejercicios

10.67 Se sabe que el contenido de los envases de un lubricante específico se distribuye normalmente con una varianza de 0.03 litros. Pruebe la hipótesis de que $\sigma^2 = 0.03$ contra la alternativa de que $\sigma^2 \neq 0.03$ para la muestra aleatoria de 10 envases del ejercicio 10.23 de la página 356. Use un valor *P* en sus conclusiones.

10.68 Por experiencia se sabe que el tiempo que se requiere para que los estudiantes de preparatoria de último año contesten una prueba estandarizada es una variable aleatoria normal con una desviación estándar de 6 minutos. Pruebe la hipótesis de que $\sigma = 6$ contra la alternativa de que $\sigma < 6$ si una muestra aleatoria de los tiempos para realizar la prueba de 20 estudiantes de preparatoria de último año tiene una desviación estándar $s = 4.51$. Utilice un nivel de significancia de 0.05.

10.69 Se deben supervisar las aflotoxinas ocasionadas por moho en cosechas de cacahuate en Virginia. Una muestra de 64 lotes de cacahuate revela niveles de 24.17 ppm, en promedio, con una varianza de 4.25 ppm. Pruebe la hipótesis de que $\sigma^2 = 4.2$ ppm contra la alternativa de que $\sigma^2 \neq 4.2$ ppm. Utilice un valor *P* en sus conclusiones.

10.70 Datos históricos indican que la cantidad de dinero que aportaron los residentes trabajadores de una ciudad grande para un escuadrón de rescate voluntario es una variable aleatoria normal con una desviación estándar de $1.40. Se sugiere que las contribuciones al escuadrón de rescate sólo de los empleados del departamento de sanidad son mucho más variables. Si las contribuciones de una muestra aleatoria de 12 empleados del departamento de sanidad tienen una desviación estándar de $1.75, ¿podemos concluir a un nivel de significancia de 0.01 que la desviación estándar de las contribuciones de todos los trabajadores de sanidad es mayor que la de todos los trabajadores que viven en dicha ciudad?

10.71 Se dice que una máquina despachadora de bebida gaseosa está fuera de control si la varianza de los contenidos excede a 1.15 decilitros. Si una muestra aleatoria de 25 bebidas de esta máquina tiene una varianza de 2.03 decilitros, ¿esto indica, a un nivel de significancia de 0.05, que la máquina está fuera de control? Suponga que los contenidos se distribuyen de forma aproximadamente normal.

10.72 Prueba de $\sigma^2 = \sigma_0^2$ para una muestra grande: Cuando $n \geq 30$ podemos probar la hipótesis nula de que $\sigma^2 = \sigma_0^2$ o $\sigma = \sigma_0$ calculando

$$z = \frac{s - \sigma_0}{\sigma_0/\sqrt{2n}},$$

que es un valor de una variable aleatoria cuya distribución muestral es aproximadamente la distribución normal estándar.

a) Con referencia al ejemplo 10.4, a un nivel de significancia de 0.05, pruebe si $\sigma = 10.0$ años contra la alternativa de que $\sigma \neq 10.0$ años.

b) Se sospecha que la varianza de la distribución de las distancias en kilómetros que un modelo nuevo de automóvil equipado con un motor diesel recorre con 5 litros de combustible es menor que la varianza de la distribución de distancias que recorre el mismo modelo equipado con un motor de gasolina de 6 cilindros, la cual se sabe es $\sigma^2 = 6.25$. Si 72 recorridos de prueba con el modelo diesel tienen una varianza de 4.41, ¿podemos concluir, a un nivel de significancia de 0.05, que la varianza de las distancias recorridas por el modelo que funciona con diesel es menor que la del modelo que funciona con gasolina?

10.73 Se realiza un estudio para comparar el tiempo que les toma a hombres y mujeres ensamblar cierto producto. La experiencia indica que la distribución del tiempo tanto para hombres como para mujeres es aproximadamente normal, pero que la varianza del tiempo para las mujeres es menor que para los hombres. Una muestra aleatoria de los tiempos de 11 hombres y 14 mujeres produce los siguientes datos:

Hombres	**Mujeres**
$n_1 = 11$	$n_2 = 14$
$s_1 = 6.1$	$s_2 = 5.3$

Pruebe la hipótesis de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 > \sigma_2^2$. Utilice un valor P en su conclusión.

10.74 En el ejercicio 10.41 de la página 358 pruebe la hipótesis a un nivel de significancia de 0.05 de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 \neq \sigma_2^2$, donde σ_1^2 y σ_2^2 son las varianzas para el número de organismos por metro cuadrado de agua en los dos lugares diferentes de Cedar Run.

10.75 Remítase al ejercicio 10.39 de la página 358 y pruebe la hipótesis de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 \neq \sigma_2^2$, donde σ_1^2 y σ_2^2 son las varianzas para la duración de las películas producidas por la empresa 1 y la empresa 2, respectivamente. Utilice un valor P.

10.76 Se comparan dos tipos de instrumentos para medir la cantidad de monóxido de azufre en la atmósfera en un experimento sobre la contaminación del aire. Los investigadores desean determinar si los dos tipos de instrumentos proporcionan mediciones con la misma variabilidad. Se registran las siguientes lecturas para los dos instrumentos:

Monóxido de azufre	
Instrumento *A*	**Instrumento *B***
0.86	0.87
0.82	0.74
0.75	0.63
0.61	0.55
0.89	0.76
0.64	0.70
0.81	0.69
0.68	0.57
0.65	0.53

Suponga que las poblaciones de mediciones se distribuyen de forma aproximadamente normal y pruebe la hipótesis de que $\sigma_A = \sigma_B$, contra la alternativa de que $\sigma_A \neq \sigma_B$. Use un valor P.

10.77 Se lleva a cabo un experimento para comparar el contenido de alcohol en una salsa de soya en dos líneas de producción diferentes. La producción se supervisa ocho veces al día. A continuación se presentan los datos.

```
Línea de producción 1.
  0.48 0.39 0.42 0.52 0.40 0.48 0.52 0.52
Línea de producción 2.
  0.38 0.37 0.39 0.41 0.38 0.39 0.40 0.39
```

Suponga que ambas poblaciones son normales. Se sospecha que la línea de producción 1 no está produciendo tan consistentemente como la línea 2 en términos de contenido de alcohol. Pruebe la hipótesis de que $\sigma_1 = \sigma_2$ contra la alternativa de que $\sigma_1 \neq \sigma_2$. Utilice un valor P.

10.78 Se sabe que las emisiones de hidrocarburos de los automóviles disminuyeron de forma drástica durante la década de 1980. Se realizó un estudio para comparar las emisiones de hidrocarburos a velocidad estacionaria, en partes por millón (ppm), para automóviles de 1980 y 1990. Se seleccionaron al azar 20 automóviles de cada modelo y se registraron sus niveles de emisión de hidrocarburos. Los datos son los siguientes:

```
Modelos 1980:
141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380
Modelos 1990:
140 160 20 20 223 60 20 95 360 70
220 400 217 58 235 380 200 175 85 65
```

Pruebe la hipótesis de que $\sigma_1 = \sigma_2$ contra la alternativa de que $\sigma_1 \neq \sigma_2$. Suponga que ambas poblaciones son normales. Utilice un valor P.

10.11 Prueba de la bondad de ajuste

A lo largo de este capítulo nos ocupamos de la prueba de hipótesis estadística acerca de parámetros de una sola población, como μ, σ^2 y p. Ahora consideraremos una prueba para determinar si una población tiene una distribución teórica específica. La prueba se basa en el nivel de ajuste que existe entre la frecuencia de ocurrencia de las observaciones en una muestra observada y las frecuencias esperadas que se obtienen a partir de la distribución hipotética.

Para ilustrar lo anterior considere el lanzamiento de un dado. Suponemos que se trata de un dado legal, lo cual equivale a probar la hipótesis de que la distribución de resultados es la distribución uniforme discreta

$$f(x) = \frac{1}{6}, \qquad x = 1, 2, \ldots, 6.$$

Suponga que el dado se lanza 120 veces y que se registra cada resultado. Teóricamente, si el dado está balanceado, esperaríamos que cada cara ocurriera 20 veces. Los resultados se presentan en la tabla 10.4.

Tabla 10.4: Frecuencias observadas y esperadas de 120 lanzamientos de un dado

Cara	**1**	**2**	**3**	**4**	**5**	**6**
Observadas	20	22	17	18	19	24
Esperadas	20	20	20	20	20	20

Al comparar las frecuencias observadas con las frecuencias esperadas correspondientes debemos decidir si es posible que tales discrepancias ocurran como resultado de fluctuaciones del muestreo, de que el dado está balanceado o no es legal o de que la distribución de resultados no es uniforme. Es práctica común referirse a cada resultado posible de un experimento como una celda. En nuestro caso tenemos 6 celdas. A continuación se define el estadístico adecuado en el cual basamos nuestro criterio de decisión para un experimento que incluye k celdas.

Una **prueba de la bondad de ajuste** entre las frecuencias observadas y esperadas se basa en la cantidad.

Prueba de la bondad de ajuste

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

donde χ^2 es un valor de una variable aleatoria cuya distribución muestral se aproxima muy de cerca a la distribución chi cuadrada con $v = k - 1$ grados de libertad. Los símbolos o_i y e_i representan las frecuencias observada y esperada, respectivamente, para la i-ésima celda.

El número de grados de libertad asociado con la distribución chi cuadrada que se utiliza aquí es igual a $k - 1$, pues sólo hay $k - 1$ frecuencias de celdas libremente determinadas. Es decir, una vez que se determinan las frecuencias de $k - 1$ celdas, también se determina la frecuencia para la k-ésima celda.

Si las frecuencias observadas se acercan a las frecuencias esperadas correspondientes, el valor χ^2 será pequeño, lo cual indica un buen ajuste. Si las frecuencias observadas difieren de manera considerable de las frecuencias esperadas, el valor χ^2 será grande y el ajuste deficiente. Un buen ajuste conduce a la aceptación de H_0, mientras que un ajuste

deficiente conduce a su rechazo. Por lo tanto, la región crítica caerá en la cola derecha de la distribución chi cuadrada. Para un nivel de significancia igual a α encontramos el valor crítico χ^2_α de la tabla A.5 y, entonces, $\chi^2 > \chi^2_\alpha$ constituye la región crítica. **El criterio de decisión que aquí se describe no se debería utilizar a menos que cada una de las frecuencias esperadas sea por lo menos igual a 5**. Esta restricción podría requerir la combinación de celdas adyacentes, lo que dará como resultado una reducción en el número de grados de libertad.

En la tabla 10.4 encontramos que el valor χ^2 es

$$\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7.$$

Si usamos la tabla A.5, encontramos $\chi^2_{0.05} = 11.070$ para $v = 5$ grados de libertad. Como 1.7 es menor que el valor crítico, no se rechaza H_0. Concluimos que no hay suficiente evidencia de que el dado está desbalanceado.

Como un segundo ejemplo probemos la hipótesis de que la distribución de frecuencias de la duración de baterías presentadas en la tabla 1.7 de la página 23 se puede aproximar mediante una distribución normal con media $\mu = 3.5$ y desviación estándar $\sigma = 0.7$. Las frecuencias esperadas para las 7 clases (celdas) que se listan en la tabla 10.5 se obtienen calculando las áreas bajo la curva normal hipotética que caen entre los diversos límites de clase.

Tabla 10.5: Frecuencias observadas y esperadas para la duración de las baterías suponiendo normalidad

Límites de clase	o_i		e_i	
1.45 – 1.95	2	} 7	0.5	} 8.5
1.95 – 2.45	1		2.1	
2.45 – 2.95	4		5.9	
2.95 – 3.45	15		10.3	
3.45 – 3.95	10		10.7	
3.95 – 4.45	5	} 8	7.0	} 10.5
4.45 – 4.95	3		3.5	

Por ejemplo, los valores z que corresponden a los límites de la cuarta clase son

$$z_1 = \frac{2.95 - 3.5}{0.7} = -0.79 \quad \text{y} \quad z_2 = \frac{3.45 - 3.5}{0.7} = -0.07.$$

En la tabla A.3 encontramos que el área entre $z_1 = -0.79$ y $z_2 = -0.07$ es

$$\begin{aligned} \text{área} = P(-0.79 < Z < -0.07) &= P(Z < -0.07) - P(Z < -0.79) \\ &= 0.4721 - 0.2148 = 0.2573. \end{aligned}$$

Por lo tanto, la frecuencia esperada para la cuarta clase es

$$e_4 = (0.2573)(40) = 10.3.$$

Se acostumbra redondear estas frecuencias a un decimal.

La frecuencia esperada para el primer intervalo de clase se obtiene utilizando el área total bajo la curva normal a la izquierda del límite 1.95. Para el último intervalo de clase usamos el área total a la derecha del límite 4.45. Todas las demás frecuencias esperadas se determinan utilizando el método que se describe para la cuarta clase. Observe que combinamos clases adyacentes en la tabla 10.5 donde las frecuencias esperadas son menores que 5 (una regla general en la prueba de la bondad de ajuste). En consecuencia, el número total de intervalos se reduce de 7 a 4, lo cual da como resultado $v = 3$ grados de libertad. Entonces, el valor χ^2 es dado por

$$\chi^2 = \frac{(7 - 8.5)^2}{8.5} + \frac{(15 - 10.3)^2}{10.3} + \frac{(10 - 10.7)^2}{10.7} + \frac{(8 - 10.5)^2}{10.5} = 3.05.$$

Como el valor χ^2 calculado es menor que $\chi^2_{0.05} = 7.815$ para 3 grados de libertad, no tenemos razón para rechazar la hipótesis nula y concluimos que la distribución normal con $\mu = 3.5$ y $\sigma = 0.7$ proporciona un buen ajuste para la distribución de la duración de las baterías.

La prueba de bondad de ajuste chi cuadrada es un recurso importante, en particular debido a que muchos procedimientos estadísticos en la práctica dependen, en un sentido teórico, de la suposición de que los datos reunidos provienen de un tipo de distribución específico. Como ya se expuso, la suposición de normalidad se hace muy a menudo. En los siguientes capítulos continuaremos haciendo suposiciones de normalidad con el fin de proporcionar una base teórica para ciertas pruebas e intervalos de confianza.

En la literatura hay pruebas para evaluar la normalidad que son más poderosas que la prueba chi cuadrada. Una de tales pruebas es la **prueba de Geary**, la cual se basa en un estadístico muy sencillo que es el cociente de dos estimadores de la desviación estándar de la población σ. Suponga que se toma una muestra aleatoria $X_1, X_2, ..., X_n$ de una distribución normal, $N(\mu, \sigma)$. Considere el cociente

$$U = \frac{\sqrt{\pi/2}\sum_{i=1}^{n} |X_i - \bar{X}|/n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}.$$

El lector debería reconocer que el denominador es un estimador razonable de σ sin importar si la distribución es normal o no. El numerador es un buen estimador de σ si la distribución es normal, pero podría sobrestimar o subestimar a σ cuando haya desviaciones de la normalidad. Así, los valores de U que difieren considerablemente de 1.0 representan la señal de que se debe rechazar la hipótesis de normalidad.

Para muestras grandes una prueba razonable se basa en la normalidad aproximada de U. El estadístico de prueba es, entonces, una estandarización de U dada por

$$Z = \frac{U - 1}{0.2661/\sqrt{n}}.$$

Desde luego, el procedimiento de prueba incluye la región crítica bilateral. Calculamos un valor de z a partir de los datos y no rechazamos la hipótesis de normalidad cuando

$$-z_{\alpha/2} < Z < z_{\alpha/2}.$$

En la bibliografía se cita un artículo que trata sobre la prueba de Geary (Geary, 1947).

10.12 Prueba de independencia (datos categóricos)

El procedimiento de prueba de chi cuadrada que se presentó en la sección 10.11 también se puede usar para probar la hipótesis de independencia de dos variables de clasificación. Suponga que deseamos determinar si las opiniones de los votantes residentes del estado de Illinois respecto a una nueva reforma fiscal son independientes de sus niveles de ingreso. Los sujetos de una muestra aleatoria de 1000 votantes registrados del estado de Illinois se clasifican de acuerdo con su posición en las categorías de ingreso bajo, medio o alto, y si están a favor o no de la nueva reforma fiscal. Las frecuencias observadas se presentan en la tabla 10.6, la cual se conoce como **tabla de contingencia**.

Tabla 10.6: Tabla de contingencia 2 × 3

	Nivel de ingreso			
Reforma fiscal	**Bajo**	**Medio**	**Alto**	**Total**
A favor	182	213	203	598
En contra	154	138	110	402
Total	336	351	313	1000

Una tabla de contingencia con r renglones y c columnas se denomina tabla $r \times c$ ("$r \times c$" se lee "r por c"). Los totales de renglones y columnas en la tabla 10.6 se denominan **frecuencias marginales**. Nuestra decisión de aceptar o rechazar la hipótesis nula, H_0, de que la opinión de un votante respecto a la nueva reforma fiscal es independiente de su nivel de ingreso, se basa en qué tan bien se ajusten las frecuencias observadas en cada una de las 6 celdas de la tabla 10.6 y en las frecuencias que esperaríamos para cada celda si supusiéramos que H_0 es verdadera. Para encontrar estas frecuencias esperadas definamos los siguientes eventos:

L: Una persona seleccionada está en el nivel de ingresos bajo.
M: Una persona seleccionada está en el nivel de ingresos medio.
H: Una persona seleccionada está en el nivel de ingresos alto.
F: Una persona seleccionada está a favor de la nueva reforma fiscal.
A: Una persona seleccionada está en contra de la nueva reforma fiscal.

Podemos usar las frecuencias marginales para listar las siguientes estimaciones de probabilidad:

$$P(L) = \frac{336}{1000}, \qquad P(M) = \frac{351}{1000}, \qquad P(H) = \frac{313}{1000},$$
$$P(F) = \frac{598}{1000}, \qquad P(A) = \frac{402}{1000}.$$

Ahora bien, si H_0 es verdadera y las dos variables son independientes, deberíamos tener

$$P(L \cap F) = P(L)\,P(F) = \left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(L \cap A) = P(L)\,P(A) = \left(\frac{336}{1000}\right)\left(\frac{402}{1000}\right),$$

$$P(M \cap F) = P(M)\,P(F) = \left(\frac{351}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(M \cap A) = P(M)\,P(A) = \left(\frac{351}{1000}\right)\left(\frac{402}{1000}\right),$$
$$P(H \cap F) = P(H)\,P(F) = \left(\frac{313}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(H \cap A) = P(H)\,P(A) = \left(\frac{313}{1000}\right)\left(\frac{402}{1000}\right).$$

Las frecuencias esperadas se obtienen multiplicando la probabilidad de cada celda por el número total de observaciones. Como antes, redondeamos estas frecuencias a un decimal. Así, se estima que el número esperado de votantes de bajo ingreso en nuestra muestra que favorecen la reforma fiscal es

$$\left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right)(1000) = \frac{(336)(598)}{1000} = 200.9$$

cuando H_0 es verdadera. La regla general para obtener la frecuencia esperada de cualquier celda es dada por la siguiente fórmula:

$$\text{frecuencia esperada} = \frac{(\text{total por columna}) \times (\text{total por renglón})}{\text{gran total}}$$

En la tabla 10.7 la frecuencia esperada para cada celda se registra entre paréntesis, a un lado del valor observado verdadero. Observe que las frecuencias esperadas en cualquier renglón o columna se suman al total marginal apropiado. En nuestro ejemplo necesitamos calcular sólo las dos frecuencias esperadas en el renglón superior de la tabla 10.7 y luego calcular las otras mediante sustracción. El número de grados de libertad asociados con la prueba chi cuadrada que aquí se usa es igual al número de frecuencias de celdas que se pueden llenar libremente cuando se nos proporcionan los totales marginales y el gran total, y en este caso ese número es 2. Una fórmula sencilla que proporciona el número correcto de grados de libertad es

$$v = (r - 1)(c - 1).$$

Tabla 10.7: Frecuencias observadas y esperadas

	Nivel de ingreso			
Reforma fiscal	**Bajo**	**Medio**	**Alto**	**Total**
A favor	182 (200.9)	213 (209.9)	203 (187.2)	598
En contra	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Por lo tanto, para nuestro ejemplo $v = (2 - 1)(3 - 1) = 2$ grados de libertad. Para probar la hipótesis nula de independencia usamos el siguiente criterio de decisión:

Prueba de independencia

Calcule

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

donde la sumatoria se extiende a todas las celdas rc en la tabla de contingencia $r \times c$.

Si $\chi^2 > \chi^2_\alpha$ con $v = (r-1)(c-1)$ grados de libertad, rechace la hipótesis nula de independencia al nivel de significancia α; en otro caso no la rechace.

Al aplicar este criterio a nuestro ejemplo encontramos que

$$\chi^2 = \frac{(182-200.9)^2}{200.9} + \frac{(213-209.9)^2}{209.9} + \frac{(203-187.2)^2}{187.2} + \frac{(154-135.1)^2}{135.1} + \frac{(138-141.1)^2}{141.1} + \frac{(110-125.8)^2}{125.8} = 7.85,$$

$$P \approx 0.02.$$

En la tabla A.5 encontramos que $\chi^2_{0.05} = 5.991$ para $v = (2-1)(3-1) = 2$ grados de libertad. Rechazamos la hipótesis nula y concluimos que la opinión de un votante respecto a la reforma fiscal y su nivel de ingresos no son independientes.

Es importante recordar que el estadístico sobre el cual basamos nuestra decisión tiene una distribución que sólo se aproxima por la distribución chi cuadrada. Los valores χ^2 calculados dependen de las frecuencias de las celdas y, en consecuencia, son discretos. La distribución chi cuadrada continua parece aproximarse muy bien a la distribución de muestreo discreta de χ^2, siempre y cuando el número de grados de libertad sea mayor que 1. En una tabla de contingencia de 2×2, donde sólo tenemos 1 grado de libertad, se aplica una corrección llamada **corrección de Yates para continuidad**.

La fórmula corregida entonces se convierte en

$$\chi^2(\text{corregida}) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

Si las frecuencias de las celdas esperadas son grandes, los resultados corregidos y sin corrección son casi iguales. Cuando las frecuencias esperadas están entre 5 y 10, se debe aplicar la corrección de Yates. Para frecuencias esperadas menores que 5 se debería utilizar la prueba exacta de Fisher-Irwin. Un análisis de esta prueba se puede encontrar en *Basic Concepts of Probability and Statistics* de Hodges y Lehmann (2005; véase la bibliografía). Sin embargo, la prueba de Fisher-Irwin se puede evitar seleccionando una muestra grande.

10.13 Prueba de homogeneidad

Cuando probamos la independencia en la sección 10.12 seleccionamos una muestra aleatoria de 1000 votantes, y determinamos al azar los totales de renglón y de columna para nuestra tabla de contingencia. Otro tipo de problema para el que se aplica el método de la sección 10.12 es aquel en el cual los totales de renglón y de columna están predeterminados. Suponga, por ejemplo, que decidimos de antemano seleccionar 200 demócratas, 150 republicanos y 150 independientes entre los votantes del estado de Carolina del Norte y registrar si están a favor de una iniciativa de ley para el aborto, si están en contra o si están indecisos. Las respuestas observadas se incluyen en la tabla 10.8.

Tabla 10.8: Frecuencias observadas

	Afiliación política			
Ley para el aborto	**Demócrata**	**Republicano**	**Independiente**	**Total**
A favor	82	70	62	214
En contra	93	62	67	222
Indeciso	25	18	21	64
Total	200	150	150	500

Ahora bien, en vez de hacer una prueba de independencia, probamos la hipótesis de que las proporciones de población dentro de cada renglón son iguales. Es decir, probamos la hipótesis de que las proporciones de demócratas, republicanos e independientes que están a favor de la ley para el aborto son iguales; las proporciones de cada afiliación política contra la ley son iguales y las proporciones de cada afiliación política que están indecisos son iguales. Básicamente nos interesamos en determinar si las tres categorías de votantes son **homogéneas** en lo que se refiere a sus opiniones acerca de la iniciativa de ley para el aborto. A esta prueba se le conoce como prueba de homogeneidad.

Al suponer homogeneidad de nuevo calculamos las frecuencias esperadas de las celdas multiplicando los totales de renglón y de columna correspondientes y después dividiendo entre el gran total. Luego continuamos el análisis utilizando el mismo estadístico chi cuadrada como antes. Ilustramos este proceso en el siguiente ejemplo para los datos de la tabla 10.8.

Ejemplo 10.14: Con respecto a los datos de la tabla 10.8 pruebe la hipótesis de que las opiniones en cuanto a la propuesta de ley para el aborto son las mismas en cada afiliación política. Utilice un nivel de significancia de 0.05.

***Solución*:**

1. H_0: Para cada opinión las proporciones de demócratas, republicanos e independientes son iguales.
2. H_1: Para al menos una opinión las proporciones de demócratas, republicanos e independientes no son iguales.
3. $\alpha = 0.05$.
4. Región crítica: $\chi^2 > 9.488$ con $v = 4$ grados de libertad.
5. Cálculos: necesitamos calcular las 4 frecuencias de las celdas usando la fórmula de las frecuencias de las celdas esperadas de la página 375. Todas las demás frecuencias se obtienen mediante sustracción. Las frecuencias de las celdas observadas y esperadas se muestran en la tabla 10.9.

Tabla 10.9: Frecuencias observadas y esperadas

	Afiliación política			
Ley para el aborto	**Demócrata**	**Republicano**	**Independiente**	**Total**
A favor	82 (85.6)	70 (64.2)	62 (64.2)	214
En contra	93 (88.8)	62 (66.6)	67 (66.6)	222
Indeciso	25 (25.6)	18 (19.2)	21 (19.2)	64
Total	200	150	150	500

Así,

$$\begin{aligned}\chi^2 &= \frac{(82-85.6)^2}{85.6} + \frac{(70-64.2)^2}{64.2} + \frac{(62-64.2)^2}{64.2} \\ &\quad + \frac{(93-88.8)^2}{88.8} + \frac{(62-66.6)^2}{66.6} + \frac{(67-66.6)^2}{66.6} \\ &\quad + \frac{(25-25.6)^2}{25.6} + \frac{(18-19.2)^2}{19.2} + \frac{(21-19.2)^2}{19.2} \\ &= 1.53.\end{aligned}$$

6. Decisión: No rechazar H_0. No hay suficiente evidencia para concluir que la proporción de demócratas, republicanos e independientes difiere para cada opinión expresada. ■

Prueba para varias proporciones

El estadístico chi cuadrada para probar la homogeneidad también se puede aplicar cuando se prueba la hipótesis de que k parámetros binomiales tienen el mismo valor. Por lo tanto, se trata de una extensión de la prueba que se presentó en la sección 10.9 para determinar las diferencias entre dos proporciones a una prueba para determinar diferencias entre k proporciones. En consecuencia, nos interesamos en probar la hipótesis nula

$$H_0: \ p_1 = p_2 = \cdots = p_k$$

contra la hipótesis alternativa H_1 de que las proporciones de la población *no son todas iguales.* Para ejecutar esta prueba primero observamos muestras aleatorias independientes de tamaños $n_1, n_2, \ldots, n_k$ de las k poblaciones y ordenamos los datos en una tabla de contingencia $2 \times k$, la tabla 10.10.

Tabla 10.10: k muestras binomiales independientes

Muestra:	**1**	**2**	$\cdots$	***k***
Éxitos	x_1	x_2	$\cdots$	x_k
Fracasos	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_k - x_k$

De acuerdo con si los tamaños de las muestras aleatorias fueron predeterminados o si ocurrieron al azar, el procedimiento de prueba es idéntico a la prueba de homogeneidad o a la prueba de independencia. Por lo tanto, las frecuencias de las celdas esperadas se calculan como antes y se sustituyen junto con las frecuencias observadas en el estadístico chi cuadrada

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

con

$$v = (2-1)(k-1) = k-1$$

grados de libertad.

Al seleccionar la región crítica apropiada de la cola superior de la forma $\chi^2 > \chi^2_\alpha$ podemos llegar ahora a una decisión respecto a H_0.

Ejemplo 10.15: En un estudio sobre un taller se reúne un conjunto de datos para determinar si la proporción de artículos defectuosos producida por los trabajadores fue la misma para el turno matutino, el vespertino y el nocturno. Los datos que se reunieron se muestran en la tabla 10.11.

Tabla 10.11: Datos para el ejemplo 10.15

Turno:	**Matutino**	**Vespertino**	**Nocturno**
Defectuosos	45	55	70
No defectuosos	905	890	870

Utilice un nivel de significancia de 0.025 para determinar si la proporción de artículos defectuosos es la misma para los tres turnos.

***Solución*:** Representemos con p_1, p_2 y p_3 la proporción verdadera de artículos defectuosos para los turnos matutino, vespertino y nocturno, respectivamente.

1. H_0: $p_1 = p_2 = p_3$.
2. H_1: p_1, p_2 y p_3 no son iguales
3. $\alpha = 0.025$.
4. Región crítica: $\chi^2 > 7.378$ para $v = 2$ grados de libertad.
5. Cálculos: En correspondencia con las frecuencias observadas $o_1 = 45$ y $o_2 = 55$, encontramos

$$e_1 = \frac{(950)(170)}{2835} = 57.0 \quad \text{y} \quad e_2 = \frac{(945)(170)}{2835} = 56.7.$$

Todas las demás frecuencias esperadas se calculan restando y se incluyen en la tabla 10.12.

Tabla 10.12: Frecuencias esperadas y observadas

Turno:	**Matutino**	**Vespertino**	**Nocturno**	**Total**
Defectuosos	45 (57.0)	55 (56.7)	70 (56.3)	170
No defectuosos	905 (893.0)	890 (888.3)	870 (883.7)	2665
Total	950	945	940	2835

Ahora bien,

$$\chi^2 = \frac{(45-57.0)^2}{57.0} + \frac{(55-56.7)^2}{56.7} + \frac{(70-56.3)^2}{56.3} + \frac{(905-893.0)^2}{893.0} + \frac{(890-888.3)^2}{888.3} + \frac{(870-883.7)^2}{883.7} = 6.29,$$

$P \approx 0.04$.

6. Decisión: no rechazamos H_0 con $\alpha = 0.025$. Sin embargo, con el valor P calculado ciertamente sería riesgoso concluir que la proporción de artículos defectuosos producidos es la misma para todos los turnos. ∎

A menudo un estudio completo implica utilizar métodos estadísticos en la prueba de hipótesis, lo que se puede mostrar a los ingenieros o científicos utilizando los

dos estadísticos de prueba, junto con valores P y gráficas estadísticas. Las gráficas complementan los diagnósticos numéricos con imágenes que indican de forma intuitiva por qué resultan esos valores P, así como qué tan razonables (o no) son las suposiciones operativas.

10.14 Estudio de caso de dos muestras

En esta sección consideramos un estudio que incluye un análisis gráfico y formal detallado, junto con la impresión por computadora con comentarios y conclusiones. En un estudio del análisis de datos que realizó el personal del Centro de Consulta Estadística del Virginia Tech se compararon dos materiales diferentes, la aleación A y la aleación B, en términos de la resistencia a la rotura. La aleación B es más costosa, aunque realmente se debería adoptar si se demuestra que es más fuerte que la aleación A. También se debe tomar en cuenta la consistencia del rendimiento de las dos aleaciones.

Se seleccionaron muestras aleatorias de vigas hechas con cada aleación y la resistencia se midió en unidades de flexión de 0.001 pulgadas cuando se aplicó una fuerza fija en ambos extremos de la viga. Se utilizaron 20 especímenes para cada una de las dos aleaciones. Los datos se presentan en la tabla 10.13.

Tabla 10.13: Datos para el estudio de caso de dos muestras

Aleación *A*			**Aleación *B***		
88	82	87	75	81	80
79	85	90	77	78	81
84	88	83	86	78	77
89	80	81	84	82	78
81	85		80	80	
83	87		78	76	
82	80		83	85	
79	78		76	79	

Es importante que el ingeniero compare las dos aleaciones. Los investigadores están interesados en la resistencia y la reproducibilidad promedio, así como en determinar si hay una violación grave de la suposición de normalidad que requieren las pruebas t y F. Las figuras 10.21 y 10.22 son gráficas de cuantil-cuantil normales de las muestras de las dos aleaciones.

Al parecer no hay ninguna violación grave de la suposición de normalidad. Además, la figura 10.23 presenta dos gráficos de caja y bigote en la misma gráfica. Los gráficos de caja y bigote sugieren que no hay una diferencia apreciable en la variabilidad de la flexión para las dos aleaciones. Sin embargo, al parecer la flexión media de la aleación B es significativamente menor, lo cual sugiere (al menos gráficamente) que la aleación B es más fuerte. Las medias muestrales y las desviaciones estándar son

$$\bar{y}_A = 83.55, \quad s_A = 3.663; \qquad \bar{y}_B = 79.70, \quad s_B = 3.097.$$

La impresión del *SAS* para el PROC TTEST se muestra en la figura 10.24. La prueba F sugiere que no hay una diferencia significativa en las varianzas ($P = 0.4709$) y el estadístico t de dos muestras para probar

$$H_0: \mu_A = \mu_B,$$
$$H_1: \mu_A > \mu_B.$$

($t = 3.59$, $P = 0.0009$) rechaza H_0 a favor de H_1 y, por consiguiente, confirma lo que sugiere la información gráfica. Aquí utilizamos la prueba t que agrupa las varianzas de dos muestras a la luz de los resultados de la prueba F. Con base en este análisis la adopción de la aleación *B* sería lo adecuado.

Significancia estadística y significancia científica o para la ingeniería

Mientras que el estadístico se podría sentir muy cómodo con los resultados de la comparación entre las dos aleaciones en el estudio de caso anterior, para el ingeniero queda un dilema. El análisis demostró una mejoría estadísticamente significativa utilizando la aleación *B*. Sin embargo, ¿realmente valdrá la pena aprovechar la diferencia que se en-

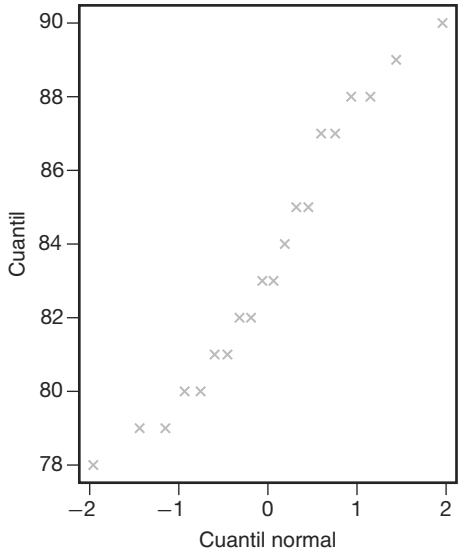


Figura 10.21: Gráfica de cuantil-cuantil normal de los datos para la aleación *A*.

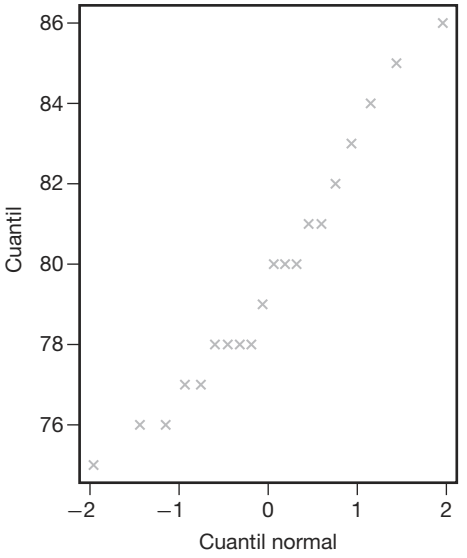


Figura 10.22: Gráfica de cuantil-cuantil normal de los datos para la aleación *B*.

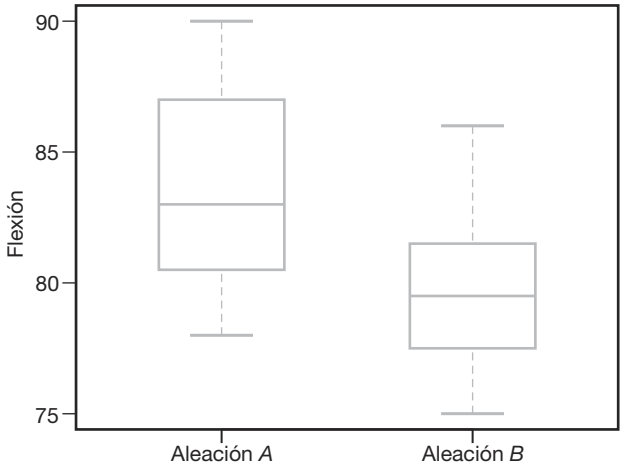


Figura 10.23: Gráficos de caja y bigote para ambas aleaciones.

contró si la aleación *B* es más costosa? Este ejemplo resalta una cuestión muy importante que con frecuencia pasan por alto los estadísticos y los analistas de datos: *la diferencia entre significancia estadística y significancia científica o para la ingeniería.* Aquí la diferencia promedio en la flexión es $\bar{y}_A - \bar{y}_B = 0.00385$ pulgadas. En un análisis completo el ingeniero debe determinar si la diferencia es suficiente para justificar el costo adicional a largo plazo. Ésta es una cuestión económica y de ingeniería. El lector debería comprender que una diferencia significativa en términos estadísticos tan sólo implica que la diferencia en las medias muestrales que se encuentra en los datos difícilmente podría ocurrir por casualidad. Esto no implica que la diferencia en las medias de la población sea profunda o particularmente significativa en el contexto del problema. Por ejemplo, en la sección 10.4 se utilizó una impresión por computadora con comentarios para demostrar la evidencia de que un medidor de pH está, de hecho, sesgado. Es decir, esto no demuestra un pH promedio de 7.00 para el material en que se probó. Pero la variabilidad entre las observaciones en la muestra es muy pequeña. El ingeniero podría decidir que las desviaciones pequeñas de 7.0 representan el medidor de pH adecuado.

```
                The TTEST Procedure
  Alloy          N    Mean  Std Dev  Std Err
Alloy A         20   83.55   3.6631   0.8191
Alloy B         20    79.7   3.0967   0.6924

  Variances     DF   t Value   Pr > |t|
  Equal         38      3.59     0.0009
  Unequal       37      3.59     0.0010
          Equality of Variances
  Num DF   Den DF   F Value   Pr > F
      19       19      1.40   0.4709
```

Figura 10.24: Impresión del *SAS* con comentarios para los datos de las aleaciones.

Ejercicios

10.79 Se supone que una máquina mezcla cacahuates, avellanas, castañas y pacanas a razón de 5:2:2:1. Se observa que una lata que contiene 500 de tales nueces mezcladas tiene 269 cacahuates, 112 avellanas, 74 castañas y 45 pacanas. A un nivel de significancia de 0.05 pruebe la hipótesis de que la máquina mezcla las nueces a una razón de 5:2:2:1.

10.80 Las calificaciones de un curso de estadística para un semestre específico fueron las siguientes:

Calificación	A	B	C	D	F
f	14	18	32	20	16

Pruebe la hipótesis, a un nivel de significancia de 0.05, de que la distribución de calificaciones es uniforme.

10.81 Se lanza un dado 180 veces con los siguientes resultados:

x	1	2	3	4	5	6
f	28	36	36	30	27	23

¿Se trata de un dado balanceado? Utilice un nivel de significancia de 0.01.

10.82 Se seleccionan tres canicas de una urna que contiene 5 canicas rojas y 3 verdes. Después de registrar el número X de canicas rojas, las canicas se reemplazan en la urna y el experimento se repite 112 veces. Los resultados que se obtienen son los siguientes:

x	0	1	2	3
f	1	31	55	25

A un nivel de significancia de 0.05, pruebe la hipótesis de que los datos registrados se pueden ajustar a la distribución hipergeométrica $h(x;\, 8, 3, 5)$, $x = 0, 1, 2, 3$.

10.83 Se lanza una moneda hasta que sale una cara y se registra el número de lanzamientos X. Después de repetir el experimento 256 veces, obtenemos los siguientes resultados:

x	1	2	3	4	5	6	7	8
f	136	60	34	12	9	1	3	1

A un nivel de significancia de 0.05, pruebe la hipótesis de que la distribución observada de X se puede ajustar a la distribución geométrica $g(x; 1/2)$, $x = 1, 2, 3,...$

10.84 En el ejercicio 1.18 de la página 31 pruebe la bondad de ajuste entre las frecuencias de clase observadas y las frecuencias esperadas correspondientes de una distribución normal con $\mu = 65$ y $\sigma = 21$. Utilice un nivel de significancia de 0.05.

10.85 En el ejercicio 1.19 de la página 31 pruebe la bondad de ajuste entre las frecuencias de clase observadas y las frecuencias esperadas correspondientes de una distribución normal con $\mu = 1.8$ y $\sigma = 0.4$. Utilice un nivel de significancia de 0.01.

10.86 En un experimento diseñado para estudiar la dependencia de la hipertensión con respecto a los hábitos de fumar se tomaron los siguientes datos de 180 individuos:

	No fumadores	**Fumadores moderados**	**Fumadores empedernidos**
Con hipertensión	21	36	30
Sin hipertensión	48	26	19

Pruebe la hipótesis de que la presencia o ausencia de hipertensión es independiente de los hábitos de tabaquismo. Utilice un nivel de significancia de 0.05.

10.87 Una muestra aleatoria de 90 adultos se clasifica de acuerdo con el género y el número de horas dedicadas a ver la televisión durante una semana:

	Género	
	Masculino	**Femenino**
Más de 25 horas	15	29
Menos de 25 horas	27	19

Utilice un nivel de significancia de 0.01 y pruebe la hipótesis de que el tiempo dedicado a ver la televisión es independiente de si el espectador es hombre o mujer.

10.88 Una muestra aleatoria de 200 hombres casados, todos jubilados, se clasificó de acuerdo con la educación y el número de hijos:

	Número de hijos		
Educación	**0–1**	**2–3**	**Más de 3**
Primaria	14	37	32
Secundaria	19	42	17
Universidad	12	17	10

Utilice un nivel de significancia de 0.05 para probar la hipótesis de que el tamaño de la familia es independiente del nivel académico del padre.

10.89 Un criminólogo realizó una investigación para determinar si la incidencia de ciertos tipos de delitos varía de una parte de una gran ciudad a otra. Los crímenes específicos de interés eran el asalto, el robo de casas, el hurto y el homicidio. La siguiente tabla muestra el número de delitos cometidos en cuatro áreas de la ciudad durante el año pasado.

	Tipo de crimen			
Distrito	**Asalto**	**Robo de casas**	**Hurto**	**Homicidio**
1	162	118	451	18
2	310	196	996	25
3	258	193	458	10
4	280	175	390	19

¿A partir de estos datos podemos concluir, a un nivel de significancia de 0.01, que la ocurrencia de estos tipos de delitos depende del distrito de la ciudad?

10.90 De acuerdo con un estudio de la Universidad Johns Hopkins, publicado en *American Journal of Public Health*, las viudas viven más que los viudos. Considere los siguientes datos reunidos de supervivencia de 100 viudas y 100 viudos después de la muerte del cónyuge:

Años vividos	**Viuda**	**Viudo**
Menos de 5	25	39
de 5 a 10	42	40
Más de 10	33	21

Con un nivel de significancia de 0.05, ¿podemos concluir que las proporciones de viudas y viudos son iguales con respecto a los diferentes periodos que un cónyuge sobrevive luego de la muerte de su compañero?

10.91 Las siguientes respuestas respecto al nivel de vida en el momento en que se aplicó una encuesta de opinión independiente a 1000 familias, comparadas con sus respuestas sobre su nivel de vida del año anterior, parecen coincidir con los resultados de un estudio publicado en *Across the Board* (junio de 1981):

	Nivel de vida			
Periodo	**Un poco mejor**	**Igual**	**No tan bueno**	**Total**
1980: Ene.	72	144	84	300
May	63	135	102	300
Sept.	47	100	53	200
1981: Ene.	40	105	55	200

Pruebe la hipótesis de que las proporciones de familias dentro de cada nivel de vida son iguales para cada uno de los cuatro periodos. Utilice un valor P.

10.92 La enfermería de una universidad realizó un experimento para determinar el grado de alivio que brindan tres jarabes para la tos. Cada jarabe se probó en 50 estudiantes y se registraron los siguientes datos:

	Jarabe para la tos		
	NyQuil	Robitussin	Triaminic
Sin alivio	11	13	9
Cierto alivio	32	28	27
Alivio completo	7	9	14

Pruebe la hipótesis de que los tres remedios para la tos son igualmente efectivos. Utilice un valor *P* en sus conclusiones.

10.93 Para determinar las posturas actuales acerca de rezar en escuelas públicas se llevó a cabo una investigación en 4 condados de Virginia. En la siguiente tabla se presentan las opiniones de 200 padres del condado de Craig, de 150 padres del condado de Giles, de 100 padres del condado de Franklin y de 100 padres del condado de Montgomery:

	Condado			
Actitud	Craig	Giles	Franklin	Mont.
A favor	65	66	40	34
En contra	42	30	33	42
Sin opinión	93	54	27	24

Pruebe la homogeneidad de las posturas entre los 4 condados respecto a rezar en escuelas públicas. Utilice un valor *P* en sus conclusiones.

10.94 Se lleva a cabo una encuesta en Indiana, Kentucky y Ohio para determinar la postura de los votantes respecto al transporte escolar. Un grupo de 200 votantes de cada uno de estos estados proporcionó los siguientes resultados:

	Postura del votante		
Estado	Apoya	No apoya	Indeciso
Indiana	82	97	21
Kentucky	107	66	27
Ohio	93	74	33

A un nivel de significancia de 0.05 pruebe la hipótesis nula de que las proporciones de votantes dentro de cada categoría de postura son las mismas en cada uno de los tres estados.

10.95 Se lleva a cabo una investigación en dos ciudades de Virginia para determinar la opinión de los votantes respecto a dos candidatos a la gubernatura en una elección próxima. En cada ciudad se seleccionaron 500 votantes al azar y se registraron los siguientes datos:

	Ciudad	
Opinión del votante	Richmond	Norfolk
A favor de *A*	204	225
A favor de *B*	211	198
Indeciso	85	77

A un nivel de significancia de 0.05 pruebe la hipótesis nula de que las proporciones de votantes que están a favor del candidato *A*, a favor del candidato *B* o que están indecisos son las mismas para cada ciudad.

10.96 En un estudio para estimar la proporción de esposas que de manera regular ven telenovelas se encuentra que 52 de 200 esposas en Denver, 31 de 150 en Phoenix y 37 de 150 en Rochester ven al menos una telenovela. Utilice un nivel de significancia de 0.05 para probar la hipótesis de que no hay diferencia entre las proporciones verdaderas de esposas que ven telenovelas en esas tres ciudades.

Ejercicios de repaso

10.97 Plantee las hipótesis nula y alternativa que utilizaría para probar las siguientes afirmaciones y determine de manera general en dónde se localiza la región crítica:

a) La cantidad promedio de nieve que cae en el lago George durante el mes de febrero es de 21.8 centímetros.
b) No más del 20% de los profesores de la universidad local contribuyó al fondo anual para donaciones.
c) En promedio, los niños asisten a la escuela en un área de 6.2 kilómetros de sus casas en un suburbio de St. Louis.
d) Al menos 70% de los automóviles nuevos del siguiente año caerán en la categoría de compactos y semicompactos.
e) La proporción de votantes que están a favor del funcionario actual para la próxima elección es de 0.58.
f) El filete rib-eye promedio en el restaurante Longhorn Steak pesa al menos 340 gramos.

10.98 Un genetista se interesa en la proporción de hombres y mujeres de una población que tiene cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres se encuentra que 31 lo padecen, mientras que sólo 24 de 100 mujeres analizadas tienen el trastorno. Con un nivel de significancia de 0.01, ¿podemos concluir que la proporción de hombres en la población con este trastorno sanguíneo es significativamente mayor que la proporción de mujeres afectadas?

10.99 Se realizó un estudio para determinar si un número mayor de italianos que de estadounidenses prefieren la champaña blanca en vez de la rosa para

las bodas. De los 300 italianos que se seleccionaron al azar, 72 preferían champaña blanca, y de los 400 estadounidenses seleccionados, 70 preferían champaña blanca en vez de la rosa. ¿Podemos concluir que una proporción mayor de italianos que de estadounidenses prefiere champaña blanca en las bodas? Utilice un nivel de significancia de 0.05.

10.100 Considere la situación del ejercicio 10.54 de la página 360. También se midió el consumo de oxígeno en mL/kg/min.

Sujeto	Con CO	Sin CO
1	26.46	25.41
2	17.46	22.53
3	16.32	16.32
4	20.19	27.48
5	19.84	24.97
6	20.65	21.77
7	28.21	28.17
8	33.94	32.02
9	29.32	28.96

Se supone que el consumo de oxígeno debería ser mayor en un ambiente relativamente libre de CO. Realice una prueba de significancia y analice la suposición.

10.101 En un estudio realizado por el Centro de Consulta Estadística de Virginia Tech se solicitó a un grupo de sujetos realizar cierta tarea en la computadora. La respuesta que se midió fue el tiempo requerido para realizar la tarea. El propósito del experimento fue probar un grupo de herramientas de ayuda desarrolladas por el Departamento de Ciencias Computacionales de la universidad. En el estudio participaron 10 sujetos. Con una asignación al azar, a 5 se les dio un procedimiento estándar usando lenguaje Fortran para realizar la tarea. A los otros 5 se les pidió realizar la tarea usando las herramientas de ayuda. A continuación se presentan los datos del tiempo requerido para completar la tarea.

Grupo 1 (procedimiento estándar)	Grupo 2 (herramienta de ayuda)
161	132
169	162
174	134
158	138
163	133

Suponga que las distribuciones de la población son normales y las varianzas son las mismas para los dos grupos y apoye o refute la conjetura de que las herramientas de ayuda aumentan la velocidad con la que se realiza la tarea.

10.102 Establezca las hipótesis nula y alternativa que usaría para probar las siguientes afirmaciones, y determine de manera general en dónde se localiza la región crítica:

a) A lo sumo, 20% de la cosecha de trigo del próximo año se exportará a la Unión Soviética.
b) En promedio, las amas de casa estadounidenses beben 3 tazas de café al día.
c) La proporción de estudiantes que se graduaron este año en Virginia, especializados en ciencias sociales, es de al menos 0.15.
d) El donativo promedio a la American Lung Association no es mayor de 10 dólares.
e) Los residentes de la zona suburbana de Richmond viajan en promedio 15 kilómetros para llegar a su lugar de trabajo.

10.103 Si se selecciona al azar una lata que contiene 500 nueces de cada uno de tres diferentes distribuidores de nueces surtidas y cada lata contiene 345, 313 y 359 cacahuates, respectivamente. Con un nivel de significancia de 0.01, ¿podríamos concluir que las nueces surtidas de los tres distribuidores contienen proporciones iguales de cacahuates?

10.104 Se realiza un estudio para determinar si hay una diferencia entre las proporciones de padres en los estados de Maryland (MD), Virginia (VA), Georgia (GA) y Alabama (AL) que están a favor de colocar Biblias en las escuelas primarias. En la siguiente tabla se registran las respuestas de 100 padres seleccionados al azar en cada uno de esos estados:

	Estado			
Preferencia	MD	VA	GA	AL
Sí	65	71	78	82
No	35	29	22	18

¿Podemos concluir que las proporciones de padres que están a favor de colocar Biblias en las escuelas son iguales en esos cuatro estados? Utilice un nivel de significancia de 0.01.

10.105 Se lleva a cabo un estudio en el Centro de Medicina Veterinaria Equina de la Universidad Regional de Virginia en Maryland para determinar si la realización de cierto tipo de cirugía en caballos jóvenes tiene algún efecto en ciertas clases de células sanguíneas del animal. Se toman muestras del fluido de seis potros antes y después de la cirugía. En las muestras se analiza el número de leucocitos de glóbulos blancos (GB) después de la operación. También se midieron los leucocitos GB preoperatorios. Los datos son los siguientes:

Potro	Precirugía*	Postcirugía*
1	10.80	10.60
2	12.90	16.60
3	9.59	17.20
4	8.81	14.00
5	12.00	10.60
6	6.07	8.60

*Todos los valores $\times 10^{-3}$.

Utilice una prueba *t* de una muestra pareada para determinar si hay un cambio significativo en los leucocitos GB con la cirugía.

10.106 El Departamento de Salud y Educación Física de Virginia Tech realizó un estudio para determinar si 8 semanas de entrenamiento realmente reducen los niveles de colesterol de los participantes. A un grupo de tratamiento que consta de 15 personas se les dieron conferencias dos veces a la semana acerca de cómo reducir sus niveles de colesterol. Otro grupo de 18 personas, de edad similar, fue seleccionado al azar como grupo de control. Se registraron los siguientes niveles de colesterol de todos los participantes al final del programa de 8 semanas:
Grupo con tratamiento:

```
Tratamiento:
    129 131 154 172 115 126 175 191
    122 238 159 156 176 175 126
Control:
    151 132 196 195 188 198 187 168 115
    165 137 208 133 217 191 193 140 146
```

¿Podemos concluir, a un nivel de significancia del 5%, que el nivel de colesterol promedio se redujo gracias al programa? Haga la prueba adecuada en las medias.

10.107 En un estudio que llevó a cabo el Departamento de Ingeniería Mecánica, el cual fue analizado por el Centro de Consulta Estadística del Virginia Tech, se compararon las varillas de acero distribuidas por dos empresas diferentes. Se fabricaron diez resortes de muestra con las varillas proporcionadas por cada empresa y se estudió la "capacidad de rebote". Los datos son los siguientes:

```
Empresa A:
9.3 8.8 6.8 8.7 8.5 6.7 8.0 6.5 9.2 7.0
Empresa B:
11.0 9.8 9.9 10.2 10.1 9.7 11.0 11.1 10.2 9.6
```

¿Puede concluir que casi no hay diferencia en las medias entre las varillas de acero proporcionadas por las dos empresas? Utilice un valor *P* para llegar a su conclusión. ¿Deberían agruparse las varianzas en este caso?

10.108 En un estudio realizado por el Centro de Recursos Acuáticos, el cual fue analizado por el Centro de Consulta Estadística del Virginia Tech, se compararon dos diferentes plantas de tratamiento para aguas residuales. La planta *A* se ubica en una zona donde el ingreso medio de los hogares está por abajo de $22,000 al año, y la planta *B* se ubica en un lugar donde el ingreso medio de los hogares está por arriba de $60,000 anuales. La cantidad de agua residual tratada en cada planta (miles de galones/día) se muestreó de forma aleatoria durante 10 días. Los datos son los siguientes:

```
Planta A:
21 19 20 23 22 28 32 19 13 18

Planta B:
20 39 24 33 30 28 30 22 33 24
```

A un nivel de significancia de 5%, ¿podemos concluir que la cantidad promedio de agua residual tratada en la planta del vecindario de altos ingresos es mayor que la tratada en la planta del área de bajos ingresos? Suponga normalidad.

10.109 Los siguientes datos muestran el número de defectos en 100,000 líneas de código en un tipo particular de software hecho en Estados Unidos y en Japón. ¿Hay suficiente evidencia para afirmar que existe una diferencia significativa entre los programas creados en los dos países? Pruebe las medias. ¿Se deberían agrupar las varianzas?

```
Estados 48 39 42 52 40 48 52 52
Unidos  54 48 52 55 43 46 48 52
Japón   50 48 42 40 43 48 50 46
        38 38 36 40 40 48 48 45
```

10.110 Existen estudios que muestran que la concentración de PCB es mucho más alta en tejido mamario maligno que en tejido mamario normal. Si un estudio de 50 mujeres con cáncer de mama revela una concentración promedio de PCB de 22.8×10^{-4} gramos, con una desviación estándar de 4.8×10^{-4} gramos, ¿la concentración media de PCB es menor que 24×10^{-4} gramos?

10.111 **Valor *z* para probar $p_1 - p_2 = d_0$.** Para probar la hipótesis nula H_0 de que $p_1 - p_2 = d_0$, donde $d_0 \neq 0$, basamos nuestra decisión en

$$z = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}},$$

que es un valor de una variable aleatoria cuya distribución se aproxima a la distribución normal estándar, siempre y cuando n_1 y n_2 sean grandes. Con respecto al ejemplo 10.11 de la página 364, pruebe la hipótesis de que el porcentaje de votantes de la ciudad que están a favor de la construcción de la planta química no excederá en más de 3% al porcentaje de votantes del condado. Utilice un valor *P* en su conclusión.

10.15 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Una de las formas más sencillas de darle un uso incorrecto a la estadística se refiere a la conclusión científica final que se obtiene cuando el analista no rechaza la hipótesis nula H_0. En este texto intentamos aclarar lo que significan la hipótesis nula y la alternativa, y también enfatizamos que, en general, la hipotesis alternativa es mucho más importante. A modo de ejemplo, si un ingeniero trata de comparar dos calibradores utilizando una prueba t de dos muestras, y H_0 afirma que "los calibradores son equivalentes", mientras que H_1 afirma que "los calibradores no son equivalentes", no rechazar H_0 no lo lleva a concluir que los calibradores son equivalentes. De hecho, ¡se puede dar el caso de que nunca se escriba o se diga "acepto H_0"! El hecho de no rechazar H_0 sólo implica que no existe evidencia suficiente. Según la naturaleza de la hipótesis, no se descartan aún muchas posibilidades.

En el capítulo 9 consideramos el caso del intervalo de confianza para muestras grandes utilizando

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

En la prueba de hipótesis es riesgoso reemplazar σ con s para $n < 30$. Si $n \geq 30$ y la distribución no es normal pero se acerca hasta cierto punto a la normal, se requiere el teorema del límite central y se confía en el hecho de que con $n \geq 30$, $s \approx \sigma$. Desde luego, cualquier prueba t va acompañada por la suposición concomitante de normalidad. Como en el caso de los intervalos de confianza, la prueba t es relativamente robusta para la normalidad. Sin embargo, cuando la muestra no es demasiado pequeña es necesario utilizar gráficas de probabilidad normal, pruebas de bondad de ajuste u otros procedimientos gráficos.

La mayoría de los capítulos de este texto incluyen análisis que tienen el propósito de relacionar el capítulo en cuestión con el siguiente material. Los temas de estimación y prueba de hipótesis se utilizan de manera importante en casi todas las técnicas que entran en el concepto de "métodos estadísticos". Los estudiantes lo notarán fácilmente cuando avancen a los capítulos 11 a 16. Será evidente que esos capítulos dependen en gran medida de los modelos estadísticos. Los estudiantes se verán expuestos al uso de los modelos en una gran variedad de aplicaciones, en diversos campos científicos y de la ingeniería. Rápidamente se darán cuenta de que el esquema de un modelo estadístico es inútil a menos que se disponga de datos para estimar parámetros en el modelo formulado. Esto será especialmente evidente en los capítulos 11 y 12, cuando se presente el concepto de modelos de regresión. Seguiremos utilizando los conceptos y la teoría relacionados con el capítulo 9. En lo que se refiere al material de este capítulo, el esquema de la prueba de hipótesis, de los valores P, de la potencia de una prueba y la selección del tamaño de la muestra, en conjunto desempeñarán un papel importante. Dado que con mucha frecuencia la formulación del modelo inicial debe complementarse con la edición del mismo antes de que el analista se sienta lo suficientemente cómodo para utilizarlo con el fin de conocer o predecir un proceso, en los capítulos 11, 12 y 15 se utilizará con frecuencia la prueba de hipótesis para complementar las medidas diagnósticas que se emplean con el fin de evaluar la calidad del modelo.

Capítulo 11

Regresión lineal simple y correlación

11.1 Introducción a la regresión lineal

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí. Por ejemplo, en una situación industrial quizá se sepa que el contenido de alquitrán en el flujo de salida de un proceso químico está relacionado con la temperatura en la entrada. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento que permita estimar el contenido de alquitrán para varios niveles de temperatura de entrada a partir de información experimental. Desde luego, es muy probable que para muchos ejemplos concretos en los que la temperatura de entrada sea la misma, por ejemplo 130ºC, el contenido de alquitrán de salida no sea el mismo. Esto es muy similar a lo que ocurre cuando se estudian varios automóviles con un motor del mismo volumen; no todos tienen el mismo rendimiento de combustible. No todas las casas ubicadas en la misma zona del país, con la misma superficie de construcción, se venden al mismo precio. El contenido de alquitrán, el rendimiento del combustible (en millas por galón) y el precio de las casas (en miles de dólares) son **variables dependientes** naturales o respuestas en los tres escenarios. La temperatura en la entrada, el volumen del motor (pies cúbicos) y los metros cuadrados de superficie de construcción son, respectivamente, **variables independientes** naturales o **regresores**. Una forma razonable de relación entre la **respuesta** Y y el regresor x es la relación lineal,

$$Y = \beta_0 + \beta_1 x,$$

en la que, por supuesto, β_0 es la **intersección** y β_1 es la **pendiente**. Esta relación se ilustra en la figura 11.1.

Si la relación es exacta y no contiene ningún componente aleatorio o probabilístico, entonces se trata de una relación **determinista** entre dos variables científicas. Sin embargo, en los ejemplos que se mencionaron, así como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista, es decir, una x dada no siempre produce el mismo valor de Y. Como resultado, los problemas importantes en este caso son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de **análisis de regresión** se refiere a encontrar la mejor relación entre Y y x

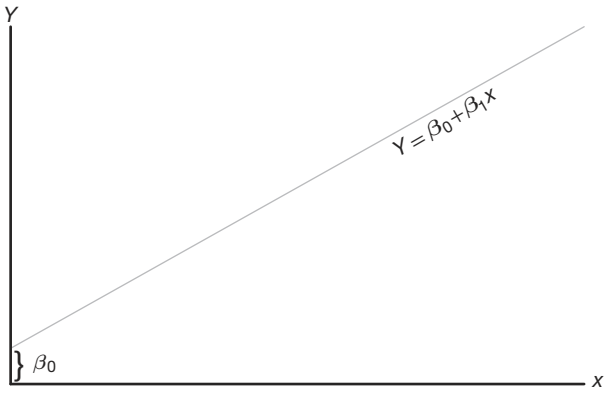


Figura 11.1: Una relación lineal; β_0: intersección; β_1: pendiente.

cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x.

En muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente **que ayude a explicar a *Y***. Por ejemplo, si se tratara de explicar las razones para el precio de una casa, se esperaría que una de ellas fuera su antigüedad, en cuyo caso la estructura múltiple de la regresión se podría escribir como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

donde Y es el precio, x_1 son los metros cuadrados y x_2 es la antigüedad de la casa en años. En el capítulo siguiente se estudiarán problemas con regresores múltiples. El análisis resultante se denomina **regresión múltiple**; en tanto que el análisis del caso con un solo regresor recibe el nombre de **regresión simple**. En un segundo ejemplo de la regresión múltiple, un ingeniero químico podría estar interesado en la cantidad de hidrógeno que se ha perdido en las muestras de un metal específico que se tiene almacenado. En este caso habría dos entradas, x_1, el tiempo de almacenamiento en horas, y x_2, la temperatura de almacenamiento en grados centígrados. De modo que la respuesta sería Y, la pérdida de hidrógeno en partes por millón.

En este capítulo estudiaremos el tema de la **regresión lineal simple**, que trata el caso de una sola variable regresora, en el que la relación entre x y y es lineal. Para el caso en el que hay más de una variable regresora el lector debe consultar el capítulo 12. Denotemos una muestra aleatoria de tamaño n mediante el conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$. Si se tomaran muestras adicionales utilizando exactamente los mismos valores de x, se esperaría que los valores de y variaran. Así, el valor y_i en el par ordenado (x_i, y_i) es el valor de cierta variable aleatoria Y_i.

11.2 El modelo de regresión lineal simple (RLS)

Hemos limitado el uso del término *análisis de regresión* a los casos en los que las relaciones entre las variables no son deterministas, es decir, no son exactas. En otras palabras, debe existir un **componente aleatorio** en la ecuación que relaciona las variables. Este componente aleatorio toma en cuenta consideraciones que no son medibles o, de

hecho, que los científicos o los ingenieros no comprenden. En realidad, en la mayoría de aplicaciones de la regresión, la ecuación lineal, digamos, $Y = \beta_0 + \beta_1 x$ es una aproximación que representa de manera simplificada algo desconocido y mucho más complicado. Por ejemplo, en el caso que implica la respuesta $Y =$ contenido de alquitrán y $x =$ temperatura de entrada es probable que $Y = \beta_0 + \beta_1 x$ sea una aproximación razonable que podría funcionar dentro de un rango limitado de x. La mayoría de las veces los modelos que son simplificaciones de estructuras más complicadas y desconocidas son de naturaleza lineal, es decir, lineales en los **parámetros** β_0 y β_1 o, en el caso del modelo que implica el precio, el tamaño y la antigüedad de la casa, lineal en los **parámetros** β_0, β_1 y β_2. Estas estructuras lineales son sencillas y de naturaleza empírica, por lo que se denominan **modelos empíricos**.

Un análisis de la relación entre x y Y requiere el planteamiento de un **modelo estadístico**. Con frecuencia un estadístico utiliza un modelo como representación de un **ideal** que, en esencia, define cómo percibimos que el sistema en cuestión generó los datos. El modelo debe incluir al conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$ de datos que implica n pares de valores (x, y). No debemos olvidar que el valor de y_i depende de x_i por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico se relaciona con la manera en que la variable aleatoria Y cambia con x y el componente aleatorio. El modelo también incluye lo que se asume acerca de las propiedades estadísticas del componente aleatorio. A continuación se presenta el modelo estadístico para la regresión lineal simple. La respuesta Y se relaciona con la variable independiente x a través de la ecuación

Modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

en la cual β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y ϵ es una variable aleatoria que se supone está distribuida con $E(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual.

En el modelo anterior hay varias cuestiones evidentes. La cantidad Y es una variable aleatoria, ya que ϵ es aleatoria. El valor x de la variable regresora no es aleatorio y, de hecho, se mide con un error despreciable. La cantidad ϵ, que a menudo recibe el nombre de **error aleatorio** o **alteración aleatoria**, tiene varianza constante. Es común que a esta parte se le denomine **suposición de varianza homogénea**. La presencia de este error aleatorio ϵ evita que el modelo se convierta tan sólo en una ecuación determinista. Ahora, el hecho de que $E(\epsilon) = 0$ implica que para una x específica, los valores de y se distribuyen alrededor de la **recta verdadera** o **recta de regresión** de la población $y = \beta_0 + \beta_1 x$. Si se elige bien el modelo, es decir, si no hay otros regresores de importancia y la aproximación lineal es buena dentro de los rangos de los datos, entonces son razonables los errores positivos y negativos que rodean a la regresión verdadera. Debe recordarse que en la práctica β_0 y β_1 se desconocen y que deben estimarse a partir de los datos. Además, el modelo que se acaba de describir es de naturaleza conceptual. Como resultado, en la práctica nunca se observan los valores ϵ reales, por lo que nunca se puede trazar la verdadera recta de regresión, aunque suponemos que ahí está. Sólo es posible dibujar una recta estimada. En la figura 11.2 se ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que sólo se dispone de $n = 5$ observaciones. Debemos destacar que lo que observamos en la figura 11.2 no es la recta que utilizan el científico o ingeniero. En vez de esa recta, ¡lo

que describe la ilustración es el significado de las suposiciones! Ahora describiremos la regresión que el usuario tiene a su disposición.

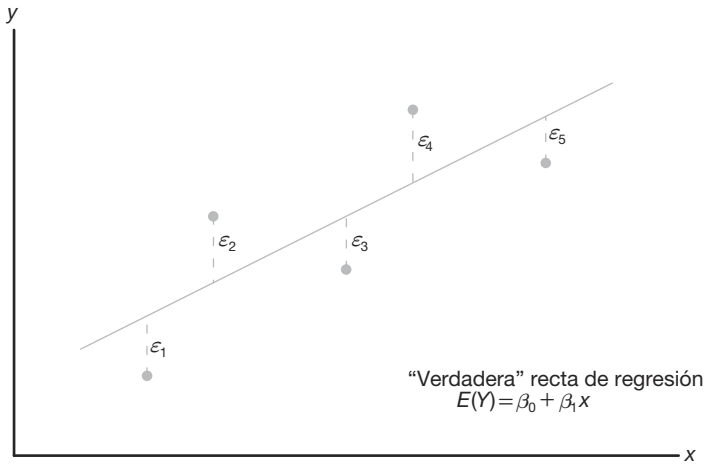


Figura 11.2: Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para $n = 5$.

La recta de regresión ajustada

Un aspecto importante del análisis de regresión es, en términos sencillos, estimar los parámetros β_0 y β_1, es decir, estimar los llamados **coeficientes de regresión**. En la sección siguiente se estudiará el método para estimarlos. Suponga que denotamos los estimados b_0 para β_0 y b_1 para β_1. Entonces, la recta de **regresión ajustada**, o estimada, es dada por

$$\hat{y} = b_0 + b_1 x,$$

donde $\hat{y}$ es el valor pronosticado o ajustado. Es evidente que la recta ajustada es un estimado de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se dispone de una gran cantidad de datos. En el ejemplo siguiente se ilustra la recta ajustada para un estudio sobre contaminación en la vida real.

Uno de los problemas más desafiantes que enfrenta el campo del control de la contaminación del agua lo representa la industria de la peletería, ya que sus desechos son químicamente complejos; se caracterizan por valores elevados de la demanda de oxígeno químico, sólidos volátiles y otras medidas de contaminación. Considere los datos experimentales de la tabla 11.1, que se obtuvieron de 33 muestras de desechos tratados químicamente en un estudio realizado en Virginia Tech. Se registraron los valores de x, la reducción porcentual de los sólidos totales, y de y, el porcentaje de disminución de la demanda de oxígeno químico.

Los datos de la tabla 11.1 aparecen graficados en un **diagrama de dispersión** en la figura 11.3. Al inspeccionar dicho diagrama se observa que los puntos se acercan mucho a una línea recta, lo cual indica que la suposición de linealidad entre las dos variables parece ser razonable.

Tabla 11.1: Medidas de la reducción de los sólidos y de la demanda de oxígeno químico

Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)	Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

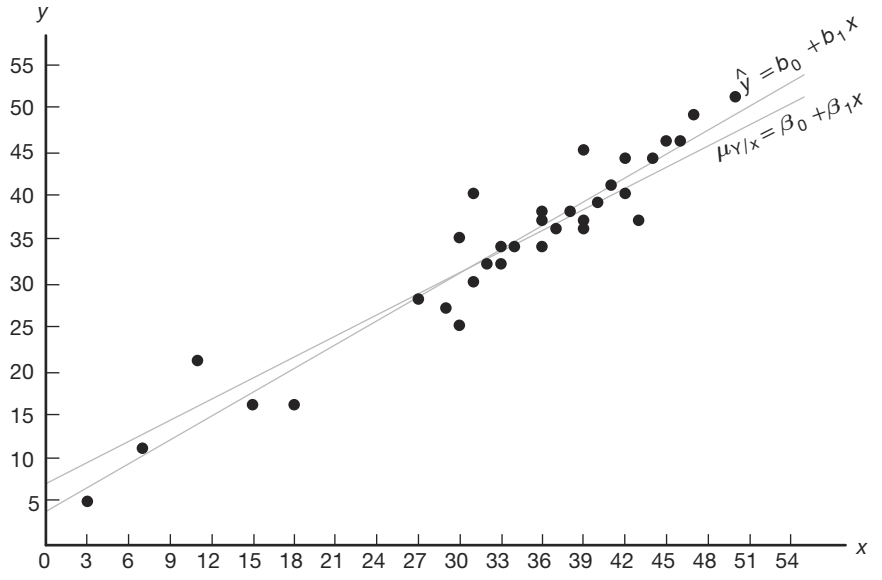


Figura 11.3: Diagrama de dispersión con rectas de regresión.

En el diagrama de dispersión de la figura 11.3 se ilustra la recta de regresión ajustada y *una recta hipotética de regresión verdadera*. Más adelante, en la sección 11.3, en la cual estudiaremos el método de estimación, revisaremos este ejemplo.

Otra mirada a las suposiciones del modelo

Resulta aleccionador repasar el modelo de regresión lineal simple que se presentó con anterioridad y analizar de forma gráfica la manera en que se relaciona con la denominada regresión verdadera. Daremos más detalles en la figura 11.2, cuando ilustremos no sólo el lugar en que los ϵ_i se localizan en la gráfica, sino también lo que implica la suposición de normalidad para los ϵ_i.

Suponga que tenemos una regresión lineal simple con $n = 6$, valores de x equidistantes y un valor único de y para cada x. Considere la gráfica de la figura 11.4, la cual debería proporcionar al lector una representación clara del modelo y de las suposiciones implicadas. La recta que aparece en la gráfica es la recta de regresión verdadera. Los puntos graficados (y, x) son puntos reales dispersos alrededor de la recta. Cada punto se ubica en su propia distribución normal, donde el centro de la distribución, es decir, la media de y, cae sobre la recta. Ciertamente esto es lo esperado, ya que $E(Y) = \beta_0 + \beta_1 x$. Como resultado, la verdadera recta de regresión **pasa a través de las medias de la respuesta** y las observaciones reales se encuentran sobre la distribución, alrededor de las medias. Observe también que todas las distribuciones tienen la misma varianza, que se denota con σ^2. Desde luego, la desviación entre una y individual y el punto sobre la recta será su valor individual ϵ. Esto queda claro porque

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i.$$

Así, con una x dada, tanto Y como el ϵ correspondiente tienen varianza σ^2.

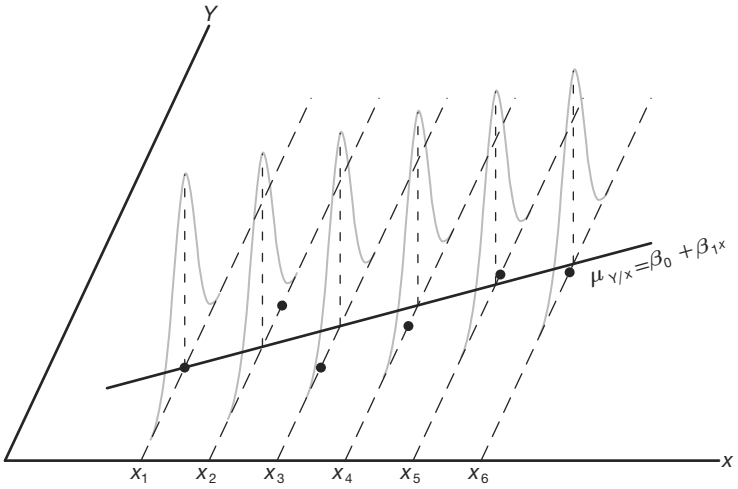


Figura 11.4: Observaciones individuales alrededor de la verdadera recta de regresión.

Note también que aquí escribimos la verdadera recta de regresión como $\mu_{Y|x} = \beta_0 + \beta_1 x$ con el fin de reafirmar que la recta pasa a través de la media de la variable aleatoria Y.

11.3 Mínimos cuadrados y el modelo ajustado

En esta sección se estudia el método para ajustar una recta de regresión estimada a los datos, lo cual equivale a determinar los estimados b_0 para β_0 y b_1 para β_1. Por supuesto,

esto permite el cálculo de los valores pronosticados a partir de la recta ajustada $\hat{y} = b_0 + b_1 x$, y otros tipos de análisis y de información diagnóstica que determinarán la fuerza de la relación, así como la adecuación y el ajuste del modelo. Antes de analizar el método de estimación de los mínimos cuadrados es importante presentar el concepto de **residual**. En esencia, un residual es un error en el ajuste del modelo $\hat{y} = b_0 + b_1 x$.

Residual: Error en el ajuste

Dado un conjunto de datos de regresión $\{(x_i, y_i); i = 1, 2,..., n\}$ y un modelo ajustado $\hat{y}_i = b_0 + b_1 x$, el i-ésimo residual e_i es dado por

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2,\ldots, n.$$

Es evidente que si un conjunto de n residuales es grande, entonces el ajuste del modelo no es bueno. Los residuales pequeños son indicadores de un ajuste adecuado. Otra relación interesante, y que a veces es útil, es la siguiente:

$$y_i = b_0 + b_1 x_i + e_i.$$

El uso de la ecuación anterior debería aclarar la diferencia entre los residuales e_i y los errores del modelo conceptual ϵ_i. No debemos olvidar que, mientras que los ϵ_i no se observan, los e_i no sólo se observan sino que desempeñan un papel importante en el análisis total.

La figura 11.5 ilustra el ajuste de la recta a este conjunto de datos: a saber $\hat{y} = b_0 + b_1 x$, y la recta que refleja el modelo $\mu_{Y|x} = \beta_0 + \beta_1 x$. Desde luego, β_0 y β_1 son parámetros desconocidos. La recta ajustada es un estimado de la recta que genera el modelo estadístico. Hay que tener presente que la recta $\mu_{Y|x} = \beta_0 + \beta_1 x$ es desconocida.

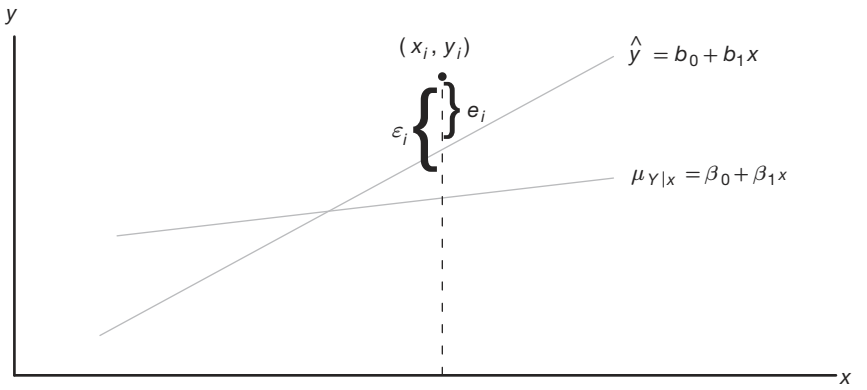


Figura 11.5: Comparación de ϵ_i con el residual e_i.

Método de mínimos cuadrados

Debemos calcular b_0 y b_1, los estimados de β_0 y β_1, de manera que la suma de los cuadrados de los residuales sea mínima. La suma residual de los cuadrados con frecuencia se denomina suma de los cuadrados del error respecto de la recta de regresión y se denota como *SCE*. Este procedimiento de minimización para estimar los parámetros

se denomina **método de mínimos cuadrados**. Por lo tanto, debemos calcular a y b para minimizar

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Al diferenciar la *SCE* con respecto a b_0 y b_1, se obtiene

$$\frac{\partial(SCE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \quad \frac{\partial(SCE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las ecuaciones siguientes (llamadas **ecuaciones normales**)

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

que se resuelven simultáneamente para obtener fórmulas de cálculo para b_0 y b_1.

Estimación de los coeficientes de regresión

Dada la muestra $\{(x_i, y_i)\}$; $i = 1, 2, \ldots, n\}$, los estimados b_0 y b_1 de los mínimos cuadrados de los coeficientes de regresión β_0 y β_1 se calculan mediante las fórmulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ y}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$

En el ejemplo siguiente se ilustra el cálculo de b_0 y b_1 usando los datos de la tabla 11.1.

Ejemplo 11.1: Estime la recta de regresión para los datos de contaminación de la tabla 11.1.

***Solución*:**

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$$

Por lo tanto,

$$b_1 = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643 \text{ y}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Por consiguiente, la recta de regresión estimada es dada por

$$\hat{y} = 3.8296 + 0.9036x.$$

Si utilizáramos la recta de regresión del ejemplo 11.1, podríamos pronosticar una reducción de 31% en la demanda de oxígeno químico si los sólidos totales se redujeran

un 30%. La reducción de 31% en la demanda de oxígeno químico se puede interpretar como un estimado de la media de la población $\mu_{Y|30}$, o como un estimado de una observación nueva si la reducción de sólidos totales es de 30%. Sin embargo, dichas estimaciones están sujetas a error. Incluso si el experimento estuviera controlado para que la reducción de los sólidos totales fuera de 30%, es improbable que la reducción en la demanda de oxígeno químico que se midiera fuera exactamente igual a 31%. De hecho, los datos originales registrados en la tabla 11.1 indican que se registraron medidas de 25% y de 35% en la reducción de la demanda de oxígeno, cuando la disminución de los sólidos totales se mantuvo en 30%.

¿Qué es lo bueno de los mínimos cuadrados?

Debemos señalar que el criterio de los mínimos cuadrados está diseñado para brindar una recta ajustada que resulte en la "cercanía" entre la recta y los puntos graficados. Existen muchas formas de medir dicha cercanía. Por ejemplo, quizá desearíamos determinar los valores de b_0 y b_1 para los que se minimiza $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ o para los que se minimiza $\sum_{i=1}^{n} |y_i - \hat{y}_i|^{1.5}$. Ambos métodos son viables y razonables. Observe que los dos, así como el procedimiento de mínimos cuadrados, obligan a que los residuales sean "pequeños" en cierto sentido. Debemos recordar que los residuales son el equivalente empírico de los valores de ϵ. La figura 11.6 ilustra un conjunto de residuales. Observe que la línea ajustada tiene valores predichos como puntos sobre la recta y, en consecuencia, los residuales son desviaciones verticales desde los puntos hasta la recta. Como resultado, el procedimiento de mínimos cuadrados genera una recta que **minimiza la suma de los cuadrados de las desviaciones verticales** desde los puntos hasta la recta.

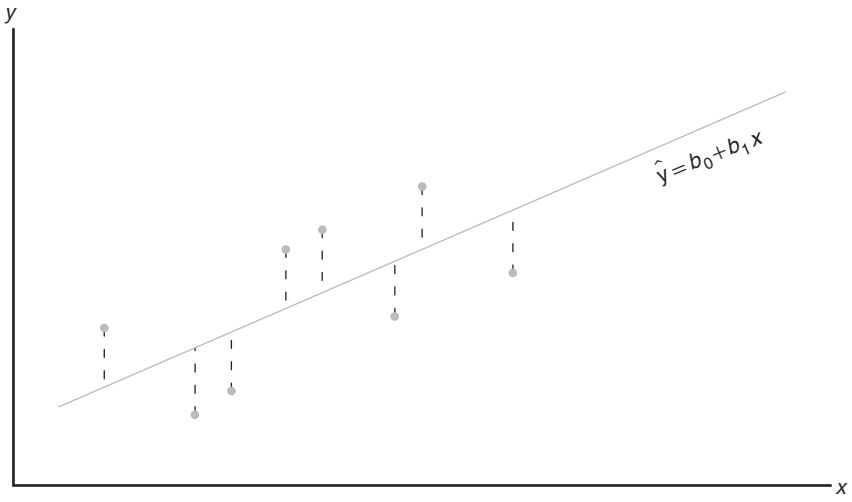


Figura 11.6: Los residuales como desviaciones verticales.

Ejercicios

11.1 Se realizó un estudio en Virginia Tech para determinar si ciertas medidas de la fuerza estática del brazo influyen en las características de "levantamiento dinámico" de un individuo. Veinticinco individuos se sometieron a pruebas de fuerza y luego se les pidió que hicieran una prueba de levantamiento de peso, en el que el peso se elevaba en forma dinámica por encima de la cabeza. A continuación se presentan los datos.

Individual	Fuerza del brazo, x	Levantamiento dinámico, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

a) Estime los valores de β_0 y β_1 para la curva de regresión lineal $\mu_{Y|x} = \beta_0 + \beta_1 x$.
b) Calcule un estimado puntual de $\mu_{Y|30}$.
c) Grafique los residuales en comparación con las x (fuerza del brazo). Comente los resultados.

11.2 Las siguientes son las calificaciones de un grupo de 9 estudiantes en un informe de medio semestre (x) y en el examen final (y):

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

a) Estime la recta de regresión lineal.
b) Calcule la calificación final de un estudiante que obtuvo 85 de calificación en el informe de medio semestre.

11.3 Se registraron las cantidades de un compuesto químico y que se disuelve en 100 gramos de agua a distintas temperaturas x:

x (°C)	y (gramos)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

a) Calcule la ecuación de la recta de regresión.
b) Grafique la recta en un diagrama de dispersión.
c) Estime la cantidad de producto químico que se disolverá en 100 gramos de agua a 50°C.

11.4 Para fines de calibración se recabaron los siguientes datos, los cuales permitirían determinar la relación entre la presión y la lectura correspondiente en la escala.

Presión, x (lb/pulg^2)	Lectura en la escala, y
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

a) Calcule la ecuación de la recta de regresión.
b) En esta aplicación el propósito de la calibración es estimar la presión a partir de una lectura observada en la escala. Estime la presión para una lectura en la escala de 54, usando $\hat{x} = (54 - b_0)/b_1$.

11.5 Se realizó un estudio sobre la cantidad de azúcar convertida en cierto proceso a distintas temperaturas. Los datos se codificaron y registraron como sigue:

Temperatura, x	Azúcar convertida, y
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

a) Estime la recta de regresión lineal.
b) Calcule la cantidad media de azúcar convertida que se produce cuando se registra una temperatura codificada de 1.75.
c) Grafique los residuales en comparación con la temperatura. Comente sus resultados.

11.6 En cierto tipo de espécimen de prueba metálico se sabe que la tensión normal sobre un espécimen se relaciona funcionalmente con la resistencia al corte. El siguiente es un conjunto de datos experimentales codificados para las dos variables:

Tensión normal, x	Resistencia al corte, y
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	24.9

a) Estime la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$.
b) Estime la resistencia al corte para una tensión normal de 24.5.

11.7 Los siguientes son algunos de los datos contenidos en un conjunto clásico denominado "datos piloto de graficación" que aparecen en *Fitting Equations to Data*, de Daniel y Wood, publicado en 1971. La respuesta y es el contenido de ácido del material determinado por análisis volumétrico; mientras que el regresor x es el contenido de ácido orgánico determinado por extracción y ponderación.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

a) Grafique los datos; ¿la regresión lineal simple parece un modelo adecuado?
b) Haga un ajuste de regresión lineal simple; calcule la pendiente y la intersección.
c) Grafique la recta de regresión en la gráfica del inciso *a*.

11.8 Se aplica un examen de colocación de matemáticas a todos los estudiantes de nuevo ingreso en una universidad pequeña. Se negará la inscripción al curso regular de matemáticas a los estudiantes que obtengan menos de 35 puntos y se les enviará a clases de regularización. Se registraron los resultados del examen de colocación y las calificaciones finales de 20 estudiantes que tomaron el curso regular:

a) Elabore un diagrama de dispersión.
b) Calcule la ecuación de la recta de regresión para predecir las calificaciones en el curso a partir de las del examen de colocación.
c) Grafique la recta en el diagrama de dispersión.
d) Si la calificación aprobatoria mínima fuera 60 puntos, ¿qué calificación en el examen de colocación se debería usar en el futuro como criterio para negar a los estudiantes el derecho de admisión a ese curso?

Examen de colocación	Calificación en el curso
50	53
35	41
35	61
40	56
55	68
65	36
35	11
60	70
90	79
35	59
90	54
80	91
60	48
60	71
60	71
40	47
55	53
50	68
65	57
50	79

11.9 Un comerciante minorista realizó un estudio para determinar la relación que hay entre los gastos semanales de publicidad y las ventas.

Costos de publicidad ($)	Ventas ($)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

a) Elabore un diagrama de dispersión.
b) Calcule la ecuación de la recta de regresión para pronosticar las ventas semanales a partir de los gastos de publicidad.
c) Estime las ventas semanales si los costos de publicidad son de $35.
d) Grafique los residuales en comparación con los costos de publicidad. Comente sus resultados.

11.10 Los siguientes datos son los precios de venta z de cierta marca y modelo de automóvil usado con w años de antigüedad. Ajuste una curva de la forma $\mu_{z|w} = \gamma\delta^w$ mediante la ecuación de regresión muestral no lineal $\hat{z} = cd^w$ [*Sugerencia:* Escriba $\ln \hat{z} = \ln c + (\ln d)w = b_0 + b_1 w$].

w (años)	z (dólares)	w (años)	z (dólares)
1	6350	3	5395
2	5695	5	4985
2	5750	5	4895

11.11 La fuerza de impulso de un motor (y) es una función de la temperatura de escape (x) en °F cuando otras variables de importancia se mantienen constantes. Considere los siguientes datos.

y	x	y	x
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

a) Grafique los datos.
b) Ajuste una recta de regresión simple a los datos y grafíquela a través de ellos.

11.12 Se realizó un estudio para analizar el efecto de la temperatura ambiente x sobre la energía eléctrica consumida por una planta química y. Otros factores se mantuvieron constantes y se recabaron los datos de una planta piloto experimental.

y (BTU)	x (°F)	y (BTU)	x (°F)
250	27	265	31
285	45	298	60
320	72	267	34
295	58	321	74

a) Grafique los datos.
b) Estime la pendiente y la intersección en un modelo de regresión lineal simple.
c) Pronostique el consumo de energía para una temperatura ambiente de 65°F.

11.13 Un estudio sobre la cantidad de lluvia y la de contaminación del aire eliminada produjo los siguientes datos:

Cantidad de lluvia diaria, x (0.01 cm)	Partículas eliminadas, y (μg/m^3)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

a) Calcule la ecuación de la recta de regresión para predecir las partículas eliminadas de la cantidad de precipitación diaria.
b) Estime la cantidad de partículas eliminadas si la precipitación diaria es $x = 4.8$ unidades.

11.14 Un profesor de la Escuela de Negocios de una universidad encuestó a una docena de colegas acerca del número de reuniones profesionales a que acudieron en los últimos cinco años (x) y el número de trabajos que enviaron a revistas especializadas (y) durante el mismo periodo. A continuación se presenta el resumen de los datos:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12,$$
$$\sum_{i=1}^{n} x_i^2 = 232, \quad \sum_{i=1}^{n} x_i y_i = 318.$$

Ajuste un modelo de regresión lineal simple entre x y y calculando los estimados de la intersección y la pendiente. Comente si la asistencia a más reuniones profesionales da como resultado más publicaciones de artículos.

11.4 Propiedades de los estimadores de mínimos cuadrados

Además de los supuestos de que el término del error en el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

es una variable aleatoria con media igual a cero y varianza σ^2 constante, suponga que además damos por hecho que ϵ_1, ϵ_2,..., ϵ_n son independientes de una corrida a otra del experimento, lo cual proporciona la base para calcular las medias y varianzas de los estimadores de β_0 y β_1.

Es importante recordar que nuestros valores de b_0 y b_1, basados en una muestra dada de n observaciones, sólo son estimaciones de los parámetros verdaderos β_0 y β_1. Si el experimento se repitiera una y otra vez, usando en cada ocasión los mismos valores fijos de x, los estimados resultantes de β_0 y β_1 muy probablemente diferirían de un experimento a otro. Estos estimados distintos podrían ser considerados como valores adoptados por las variables aleatorias B_0 y B_1; en tanto que b_0 y b_1 son ejecuciones específicas.

Como los valores de x permanecen fijos, los valores de B_0 y B_1 dependen de las variaciones en los valores de y o, con más precisión, en los valores de las variables aleatorias

Y_1, Y_2,..., Y_n. Las suposiciones sobre la distribución implican que las Y_i, $i = 1, 2,..., n$ también están distribuidas de manera independiente, con media $\mu_{Y|x_i} = \beta_0 + \beta_1 x_i$ y varianzas σ^2 iguales, es decir,

$$\sigma^2_{Y|x_i} = \sigma^2 \quad \text{para } i = 1, 2, \ldots, n.$$

Media y varianza de los estimadores

En la exposición que sigue mostramos que el estimador B_1 es insesgado para β_1, y se demuestran tanto las varianzas de B_0 como las de B_1. Esto inicia una serie de procedimientos que conducen a la prueba de hipótesis y a la estimación de intervalos de confianza para la intersección y la pendiente.

Como el estimador

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

es de la forma $\sum_{i=1}^{n} c_i Y_i$,

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad i = 1, 2, \ldots, n,$$

podemos concluir a partir del teorema 7.11 que B_1 tiene una distribución $n(\mu_{B_1}, \sigma_{B_1})$ con

$$\mu_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 \text{ y } \sigma^2_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2_{Y_i}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

También se puede demostrar (véase el ejercicio de repaso 11.60 de la página 438) que la variable aleatoria B_0 se distribuye normalmente con

$$\text{media } \mu_{B_0} = \beta_0 \text{ y varianza } \sigma^2_{B_0} = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

A partir de estos resultados es evidente que los **estimadores de mínimos cuadrados tanto para β_0 como para β_1 son insesgados**.

Partición de la variabilidad total y estimación de σ^2

Para hacer inferencias sobre β_0 y β_1 es necesario llegar a una estimación del parámetro σ^2 que aparece en las dos fórmulas anteriores de la varianza de B_0 y B_1. El parámetro σ^2, el modelo de la varianza del error, refleja una variación aleatoria o una variación del

error experimental alrededor de la recta de regresión. En gran parte de lo que sigue se recomienda emplear la notación

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

De manera que la suma de los cuadrados del error se puede escribir como sigue:

$$\begin{aligned} SCE &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}, \end{aligned}$$

que es el paso final que surge del hecho de que $b_1 = S_{xy}/S_{xx}$.

Teorema 11.1: Un estimador insesgado de σ^2 es

$$s^2 = \frac{SCE}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

La prueba del teorema 11.1 se deja como ejercicio (véase el ejercicio de repaso 11.59).

El estimador de σ^2 como error cuadrado medio

Para darnos una idea del estimador de σ^2 deberíamos observar el resultado del teorema 11.1. El parámetro σ^2 mide la varianza o las desviaciones cuadradas entre los valores de Y y su media, dada por $\mu_{Y|x}$, es decir, las desviaciones cuadradas entre Y y $\beta_0 + \beta_1 x$. Por supuesto, $\beta_0 + \beta_1 x$ se estima por medio de $\hat{y} = b_0 + b_1 x$. Por consiguiente, tendría sentido que la varianza σ^2 se describa mejor como una desviación cuadrada de la observación típica y_i con respecto a la media estimada $\hat{y}_i$, que es el punto correspondiente sobre la recta ajustada. Entonces, los valores $(y_i - \hat{y}_i)$ revelan la varianza apropiada, de manera muy similar a como los valores $(y_i - \bar{y})^2$ miden la varianza cuando se realiza un muestreo en un escenario no relacionado con la regresión. En otras palabras, $\bar{y}$ estima la media en la última situación sencilla, mientras que $\hat{y}_i$ estima la media de y_i en una estructura de regresión. Ahora, ¿qué significa el divisor $n - 2$? En las secciones que siguen observaremos que éstos son los grados de libertad asociados con el estimador s^2 de σ^2. En tanto que en el escenario i.i.d. (independiente e idénticamente distribuidas), la normal estándar se resta un grado de libertad de n en el denominador, para lo cual una explicación razonable es que se estima un parámetro, que es la media μ por medio de, digamos, $\bar{y}$, pero en el problema de la regresión **se estiman dos parámetros**, que son β_0 y β_1, por medio de b_0 y b_1. Así, el parámetro importante σ^2, que se estima mediante

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2),$$

se denomina **error cuadrado medio**, que describe un tipo de media (división entre $n - 2$) de los residuales cuadrados.

11.5 Inferencias sobre los coeficientes de regresión

Además de tan sólo estimar la relación lineal entre x y Y para fines de predicción, el experimentador podría estar interesado en hacer ciertas inferencias acerca de la pendiente y la intersección. Para dar ocasión a la prueba de hipótesis y a la construcción de intervalos de confianza para β_0 y β_1, debemos estar dispuestos a hacer la suposición adicional de que cada ϵ_i, $i = 1, 2, ..., n$, se distribuye de forma normal. Esta suposición implica que Y_1, Y_2,..., Y_n también están distribuidas normalmente, cada una con una distribución de probabilidad $n(y_i; \beta_0 + \beta_1 x_i, \sigma)$.

A partir de la sección 11.4 sabemos que B_1 tiene una distribución normal, y suponiendo normalidad, un resultado muy parecido al que se plantea en el teorema 8.4 nos permite concluir que $(n - 2)S^2/\sigma^2$ es una variable chi cuadrada con $n - 2$ grados de libertad, independiente de la variable aleatoria B_1. Entonces, el teorema 8.5 garantiza que el estadístico

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

tenga una distribución t con $n - 2$ grados de libertad. Podemos utilizar el estadístico T para construir un intervalo de confianza del $100(1 - \alpha)\%$ para el coeficiente β_1.

Intervalo de confianza para β_1

Un intervalo de confianza de $100(1 - \alpha)\%$ para el parámetro β_1 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$ es

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.2: Calcule un intervalo de confianza de 95% para β_1 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$, con base en los datos de contaminación de la tabla 11.1.

***Solución*:** A partir de los resultados dados en el ejemplo 11.1, se determina que $S_{xx} = 4152.18$ y $S_{xy} = 3752.09$. Además, se observa que $S_{yy} = 3713.88$. Recuerde que $b_1 = 0.903643$. En consecuencia,

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n - 2} = \frac{3713.88 - (0.903643)(3752.09)}{31} = 10.4299.$$

Por lo tanto, al sacar la raíz cuadrada obtenemos $s = 3.2295$. Si usamos la tabla A.4 encontramos que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Así, un intervalo de confianza de 95% para β_1 es

$$0.903643 - \frac{(2.045)(3.2295)}{\sqrt{4152.18}} < \beta < 0.903643 + \frac{(2.045)(3.2295)}{\sqrt{4152.18}},$$

que se simplifica a

$$0.8012 < \beta_1 < 1.0061.$$

Prueba de hipótesis sobre la pendiente

Para probar la hipótesis nula H_0 de que $\beta_1 = \beta_{10}$, en comparación con una alternativa posible, utilizamos de nuevo la distribución t con $n - 2$ grados de libertad con el fin de establecer una región crítica y después basar nuestra decisión en el valor de

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}.$$

El método se ilustra con el ejemplo siguiente.

Ejemplo 11.3: Utilice el valor estimado $b_1 = 0.903643$ del ejemplo 11.1 y pruebe la hipótesis de que $\beta_1 = 1.0$ en comparación con la alternativa de que $\beta_1 < 1.0$.

***Solución*:** Las hipótesis son H_0: $\beta_1 = 1.0$ y H_1: $\beta_1 < 1.0$. Por lo tanto,

$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

con $n - 2 = 31$ grados de libertad ($P \approx 0.03$).

Decisión: El valor t es significativo al nivel 0.03, lo cual sugiere evidencia sólida de que $\beta_1 < 1.0$. ∎

Una prueba t importante sobre la pendiente es la prueba de la hipótesis

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } H_1\colon\ \beta_1 \neq 0.$$

Cuando no se rechaza la hipótesis nula la conclusión es que no hay relación lineal significativa entre $E(y)$ y la variable independiente x. La gráfica de los datos del ejemplo 11.1 sugeriría que existe una relación lineal. Sin embargo, en ciertas aplicaciones en las que σ^2 es grande y, por ende, hay "ruido" considerable en los datos, una gráfica, aunque útil, quizá no produzca información clara para el investigador. El rechazo anterior de H_0 implica que hay una relación lineal significativa.

La figura 11.7 muestra la salida de resultados de *MINITAB* que presenta la prueba t para

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } H_1\colon\ \beta_1 \neq 0,$$

para los datos del ejemplo 11.1. Observe el coeficiente de regresión (Coef), el error estándar (EE Coef), el valor t (T) y el valor P (P). Se rechaza la hipótesis nula. Es claro que existe una relación lineal significativa entre la reducción de la demanda media del oxígeno químico y la reducción de los sólidos. Observe que el estadístico t se calcula como

$$t = \frac{\text{coeficiente}}{\text{error estándar}} = \frac{b_1}{s/\sqrt{S_{xx}}}.$$

El no rechazo de H_0: $\beta_1 = 0$ sugiere que no hay una relación lineal entre Y y x. La figura 11.8 es una ilustración de la implicación de este resultado; podría significar que los cambios de x tienen poco efecto sobre los cambios de Y, como se ve en el inciso *a*. Sin embargo, también puede indicar que la relación verdadera es no lineal, como se aprecia en *b*.

Cuando se rechaza H_0: $\beta_1 = 0$ existe la implicación de que el término lineal en x que reside en el modelo explica una parte significativa de la variabilidad de Y. Las dos gráfi-

```
Regression Analysis: COD versus Per_Red
The regression equation is COD = 3.83 + 0.904 Per_Red

Predictor       Coef   SE Coef       T      P
Constant       3.830     1.768    2.17  0.038
Per_Red     0.90364   0.05012   18.03  0.000

S = 3.22954    R-Sq = 91.3%    R-Sq(adj ) = 91.0%
Analysis of Variance
Source            DF       SS       MS        F      P
Regression         1   3390.6   3390.6   325.08  0.000
Residual Error    31    323.3     10.4
Total             32   3713.9
```

Figura 11.7: Salida de resultados de *MINITAB* para la prueba *t* de los datos del ejemplo 11.1.

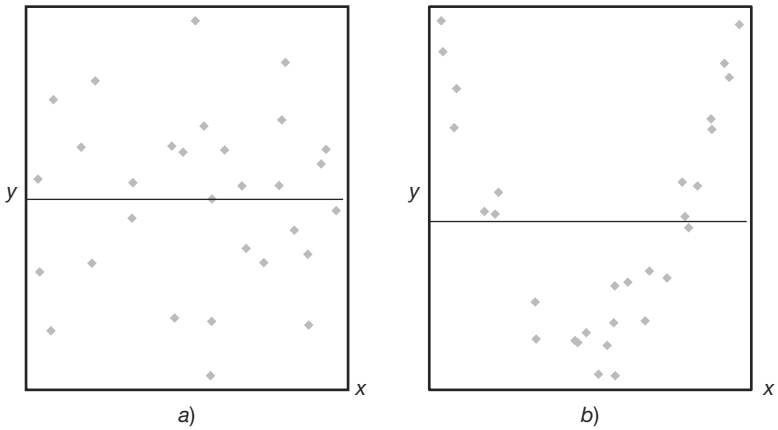


Figura 11.8: No se rechaza la hipótesis H_0: $\beta_1 = 0$.

cas que aparecen en la figura 11.9 ilustran los escenarios posibles. Como se muestra en el inciso *a* de la figura, el rechazo de H_0 sugiere que la relación en efecto es lineal. En el caso del inciso *b*, lo que se observa sugiere que, aunque el modelo contenga un efecto lineal, se podría obtener una mejor representación si se incluyera un término polinomial (tal vez cuadrático), es decir, términos que complementen el término lineal.

Inferencia estadística sobre la intersección

Los intervalos de confianza y la prueba de hipótesis del coeficiente β_0 se podrían establecer a partir del hecho de que B_0 también se distribuye de forma normal. No es difícil demostrar que

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

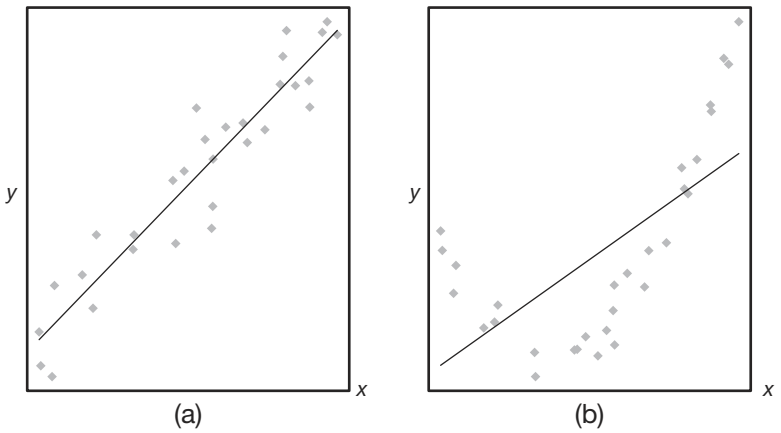


Figura 11.9: Se rechaza la hipótesis de que H_0: $\beta_1 = 0$.

tiene una distribución t con $n - 2$ grados de libertad, de manera que podemos construir un intervalo de confianza de $100(1 - \alpha)\%$ para α.

Intervalo de confianza para β_0

Un intervalo de confianza de $100(1 - \alpha)\%$ para el parámetro β_0 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$ es

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.4: Calcule un intervalo de confianza de 95% para β_0 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$ con base en los datos de la tabla 11.1.

***Solución*:** En los ejemplos 11.1 y 11.2 se encontró que

$$S_{xx} = 4152.18 \quad \text{y} \quad s = 3.2295.$$

Del ejemplo 11.1 se tiene que

$$\sum_{i=1}^{n} x_i^2 = 41{,}086 \quad \text{y} \quad b_0 = 3.829633.$$

Si usamos la tabla A.4, encontramos que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para β_0 es

$$3.829633 - \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}} < \beta_0 < 3.829633 + \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}},$$

que se simplifica a $0.2132 < \beta_0 < 7.4461$.

Para probar la hipótesis nula H_0 de que $\beta_0 = \beta_{00}$ en comparación con una alternativa posible utilizamos la distribución t con $n - 2$ grados de libertad para establecer una región crítica y, luego, basar nuestra decisión en el valor de

$$t = \frac{b_0 - \beta_{00}}{s\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}.$$

Ejemplo 11.5: Utilice el valor estimado de $b_0 = 3.829633$ del ejemplo 11.1 y, a un nivel de significancia de 0.05, pruebe la hipótesis de que $\beta_0 = 0$ en comparación con la alternativa de que $\beta_0 \neq 0$. Entonces

***Solución*:** Las hipótesis son H_0: $\beta_0 = 0$ y H_1: $\beta_0 \neq 0$. Así que,

$$t = \frac{3.829633 - 0}{3.2295\sqrt{41{,}086/[(33)(4152.18)]}} = 2.17,$$

con 31 grados de libertad. Por lo tanto, $P =$ valor $P \approx 0.038$ y concluimos que $\beta_0 \neq 0$. Observe que esto tan sólo es Coef/desviación estándar, como se aprecia en la salida de resultados de *MINITAB* en la figura 11.7. El SE Coef es el error estándar de la intersección estimada. ■

Una medida de la calidad del ajuste: el coeficiente de determinación

Observe en la figura 11.7 que aparece un elemento denotado con R-Sq, cuyo valor es 91.3%. Esta cantidad, R^2, se denomina **coeficiente de determinación** y es una medida de la **proporción de la variabilidad explicada por el modelo ajustado**. En la sección 11.8 se presentará el concepto del método del análisis de varianza para la prueba de hipótesis en la regresión. El enfoque del análisis de varianza utiliza la suma de los cuadrados del error $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ y la **suma total de los cuadrados corregida** $STCC = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$. Esta última representa la variación en los valores de respuesta que *idealmente* serían explicados con el modelo. El valor de la *SCE* es la variación debida al error, o la **variación no explicada**. Resulta claro que si la $SCE = 0$, toda variación queda explicada. La cantidad que representa la variación explicada es $STCC - SCE$. R^2 es el

$$\text{Coeficiente de determinación: } R^2 = 1 - \frac{SCE}{STCC}.$$

Advierta que si el ajuste es perfecto, *todos los residuales son cero*, y así $R^2 = 1.0$. Pero si la *SCE* es tan sólo un poco menor que la *STCC*, $R^2 \approx 0$. Observe en la salida de resultados de la figura 11.7 que el coeficiente de determinación sugiere que el modelo ajustado a los datos explica el 91.3% de la variabilidad observada en la respuesta, la reducción en la demanda de oxígeno químico.

La figura 11.10 ofrece ejemplos de una gráfica con un buen ajuste ($R^2 \approx 1.0$) en *a*) y una gráfica con un ajuste deficiente ($R^2 \approx 0$) en *b*).

Errores en el uso de R^2

Los analistas citan con mucha frecuencia los valores de R^2, quizá debido a su simplicidad. Sin embargo, hay errores en su interpretación. La confiabilidad de R^2 depende del

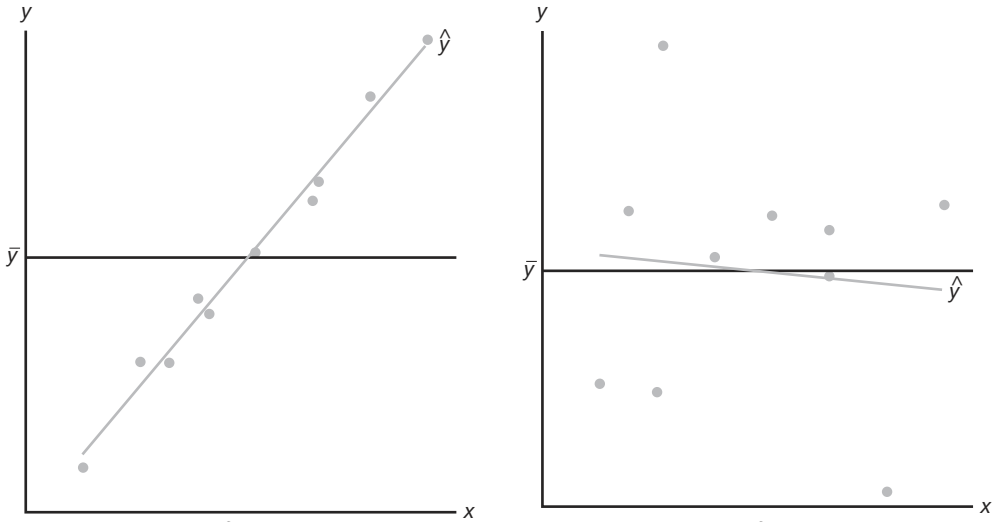


Figura 11.10: Gráficas que ilustran un ajuste muy bueno y otro deficiente.

tamaño del conjunto de los datos de la regresión y del tipo de aplicación. Resulta claro que $0 \leq R^2 \leq 1$, y el límite superior se logra cuando el ajuste a los datos es perfecto, es decir, cuando todos los residuales son cero. ¿Cuál es un valor aceptable de R^2? Se trata de una pregunta difícil de responder. Un químico encargado de establecer una calibración lineal de una pieza de equipo de alta precisión seguramente esperaría obtener un valor muy alto de R^2 (quizá superior a 0.99); mientras que un científico del comportamiento, que trabaja con datos en los que influye la variabilidad de la conducta humana, quizá se sentiría afortunado si obtuviera un valor de R^2 de hasta 0.70. Un individuo con experiencia en el ajuste de modelos tiene la sensibilidad para saber cuándo un valor es suficientemente grande dada la situación que está enfrentando. Es evidente que algunos fenómenos científicos se prestan más a un modelamiento más preciso que otros.

Es peligroso usar el criterio de R^2 para comparar *modelos en competencia* para el mismo conjunto de datos. Cuando se agregan términos adicionales al modelo, por ejemplo un regresor más, disminuye la *SCE*, lo que provoca que R^2 aumente (o al menos no disminuya). Esto implica que R^2 se puede volver artificialmente elevado por medio de la práctica inapropiada de **sobreajustar**, es decir, de incluir demasiados términos en el modelo. Por consiguiente, el incremento inevitable de R^2 que se logra al agregar términos adicionales no implica que éstos se necesitaban. En realidad, el modelo simple puede ser mejor para predecir los valores de la respuesta. En el capítulo 12, cuando se presente el concepto de los modelos que implican **más de un solo regresor**, se estudiará con detalle el papel del sobreajuste y su influencia sobre la capacidad de predicción. En este momento baste decir que *para seleccionar un modelo no se debe adoptar un proceso de selección que sólo incluya la consideración de* R^2.

11.6 Predicción

Hay varias razones para construir un modelo de regresión lineal. Una de ellas es, desde luego, predecir valores de respuesta para uno o más valores de la variable independiente. En esta sección se centra el enfoque en los errores asociados con la predicción.

La ecuación $\hat{y} = b_0$ y $b_1 x$ se puede utilizar para predecir o estimar la **respuesta media** $\mu_{Y|x_0}$ en $x = x_0$, donde x_0 no necesariamente es uno de los valores preestablecidos, o cuando $x = x_0$, se podría emplear para pronosticar un solo valor y_0 de la variable Y_0. Se esperaría que el error de predicción fuera mayor para el caso de un solo valor pronosticado que para aquel en que se predice una media. Entonces, esto afectaría la anchura de los intervalos para los valores que se predicen.

Suponga que el experimentador desea construir un intervalo de confianza para $\mu_{Y|x_0}$. En tal caso debe usar el estimador puntual $\hat{Y}_0 = B_0$ y $B_1 x_0$ para estimar $\mu_{Y|x_0} = \beta_0 + \beta_1 x$. Se puede demostrar que la distribución muestral de $\hat{Y}_0$ es normal con media

$$\mu_{Y|x_0} = E(\hat{Y}_0) = E(B_0 + B_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_{Y|x_0}$$

y varianza

$$\sigma^2_{\hat{Y}_0} = \sigma^2_{B_0 + B_1 x_0} = \sigma^2_{\bar{Y} + B_1(x_0 - \bar{x})} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right],$$

esta última surge del hecho de que $\text{Cov}(\bar{Y}_0, B_1) = 0$ (véase el ejercicio de repaso 11.61 de la página 438). Por consiguiente, ahora podemos construir un intervalo de confianza de $100(1 - \alpha)\%$ sobre la respuesta media $\mu_{Y|x_0}$ a partir del estadístico

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0 - \bar{x})^2 / S_{xx}}},$$

que tiene una distribución t con $n - 2$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_0}$

Un intervalo de confianza de $100(1 - \alpha)\%$ para la respuesta media $\mu_{Y|x_0}$ es

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

$t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.6: Con los datos de la tabla 11.1 construya límites de confianza de 95% para la respuesta media $\mu_{Y|x_0}$.

***Solución*:** A partir de la ecuación de regresión encontramos que, para $x_0 = 20\%$ de reducción de sólidos, digamos,

$$\hat{y}_0 = 3.829633 + (0.903643)(20) = 21.9025.$$

Además, $\bar{x} = 33.4545$, $S_{xx} = 4152.18$, $s = 3.2295$ y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para $\mu_{Y|20}$ es

$$21.9025 - (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < \mu_{Y|20} < 21.9025 + (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

o simplemente, $20.1071 < \mu_{Y|20} < 23.6979$.

Si repetimos los cálculos anteriores para cada uno de los diferentes valores de x_0, obtenemos los límites de confianza correspondientes para cada $\mu_{Y|x_0}$. En la figura 11.11 se presentan los datos de los puntos, la recta de regresión estimada y los límites de confianza superior e inferior sobre la media de $Y|x$.

En el ejemplo 11.6 tenemos 95% de confianza en que la reducción media poblacio-

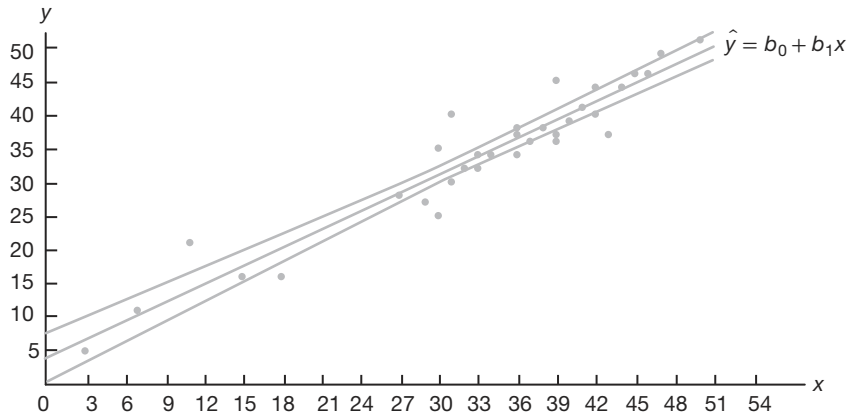


Figura 11.11: Límites de confianza para el valor medio de $Y|x$.

nal en la demanda de oxígeno químico estará entre el 20.1071% y 23.6979%, cuando la reducción de sólidos sea de 20%.

Predicción del intervalo

Otro tipo de intervalo que con frecuencia se malinterpreta y se confunde con aquel dado para $\mu_{Y|x}$ es el intervalo de la predicción para una respuesta futura observada. En realidad, en muchos casos el intervalo de la predicción es más relevante para el científico o el ingeniero que el intervalo de confianza sobre la media. En el ejemplo del contenido de alquitrán y la temperatura de entrada, mencionado en la sección 11.1, seguramente sería interesante no sólo estimar la media del contenido de alquitrán a una temperatura específica, sino también construir un intervalo que refleje el error en la predicción de una cantidad futura observada del contenido de alquitrán a la temperatura dada.

Para obtener un **intervalo de predicción** para cualquier valor único y_0 de la variable Y_0 es necesario estimar la varianza de las diferencias entre las ordenadas $\hat{y}_0$, obtenidas de las rectas de regresión calculadas en el muestreo repetido cuando $x = x_0$, y la ordenada verdadera correspondiente y_0. Podríamos considerar la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y}_0 - Y_0$, cuya distribución muestral se podría demostrar que es normal con media

$$\mu_{\hat{Y}_0 - Y_0} = E(\hat{Y}_0 - Y_0) = E[B_0 + B_1 x_0 - (\beta_0 + \beta_1 x_0 + \epsilon_0)] = 0$$

y varianza

$$\sigma^2_{\hat{Y}_0 - Y_0} = \sigma^2_{B_0 + B_1 x_0 - \epsilon_0} = \sigma^2_{\bar{Y} + B_1(x_0 - \bar{x}) - \epsilon_0} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right].$$

Así, un intervalo de predicción de $100(1-\alpha)\%$ para un solo valor pronosticado y_0 se puede construir a partir del estadístico

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

que tiene una distribución t con $n-2$ grados de libertad.

Intervalo de predicción para y_0

Un intervalo de predicción de $100(1-\alpha)\%$ para una sola respuesta y_0 es dado por

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Es claro que hay una diferencia entre el concepto de un intervalo de confianza y el del intervalo de predicción antes descrito. La interpretación del intervalo de confianza es idéntica a la que se describió para todos los intervalos de confianza sobre los parámetros de la población estudiados en el libro. De hecho, $\mu_{Y|x_0}$ es un parámetro de la población. Sin embargo, el intervalo de la predicción calculado representa un intervalo que tiene una probabilidad igual a $1-\alpha$ de contener no un parámetro sino un valor futuro de y_0 de la variable aleatoria Y_0.

Ejemplo 11.7: Con los datos de la tabla 11.1 construya un intervalo de predicción de 95% para y_0 cuando $x_0 = 20\%$.

***Solución*:** Tenemos que $n = 33$, $x_0 = 20$, $\bar{x} = 33.4545$, $\hat{y}_0 = 21.9025$, $S_{xx} = 4152.18$, $s = 3.2295$, y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de predicción de 95% para y_0 es

$$21.9025 - (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < y_0$$
$$< 21.9025 + (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

que se simplifica como $15.0585 < y_0 < 28.7464$.

En la figura 11.12 se presenta otra gráfica de los datos de reducción de la demanda de oxígeno químico, tanto con los intervalos de confianza de la respuesta media como con el intervalo de predicción sobre una respuesta individual. En el caso de la respuesta media la gráfica refleja un intervalo mucho más angosto alrededor de la recta de regresión.

Ejercicios

11.15 Remítase al ejercicio 11.1 de la página 398,
a) evalúe s^2;
b) pruebe la hipótesis de que $\beta_1 = 0$ en comparación con la alternativa de que $\beta_1 \neq 0$ a un nivel de significancia de 0.05, e interprete la decisión resultante.

11.16 Remítase al ejercicio 11.2 de la página 398,
a) evalúe s^2;
b) construya un intervalo de confianza de 95% para β_0;
c) construya un intervalo de confianza de 95% para β_1.

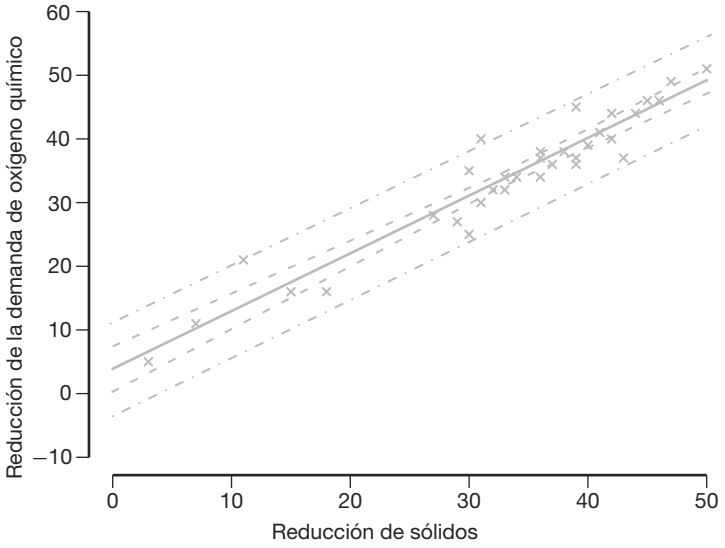


Figura 11.12: Intervalos de confianza y predicción para los datos de la reducción de la demanda de oxígeno químico; las bandas internas indican los límites de confianza para las respuestas medias y las externas señalan los límites de predicción para las respuestas futuras.

11.17 Remítase al ejercicio 11.5 de la página 398,
a) evalúe s^2;
b) construya un intervalo de confianza de 95% para β_0;
c) construya un intervalo de confianza de 95% para β_1.

11.18 Remítase al ejercicio 11.6 de la página 399,
a) evalúe s^2;
b) construya un intervalo de confianza de 99% para β_0;
c) construya un intervalo de confianza de 99% para β_1.

11.19 Remítase al ejercicio 11.3 de la página 398,
a) evalúe s^2;
b) construya un intervalo de confianza de 99% para β_0;
c) construya un intervalo de confianza de 99% para β_1.

11.20 Pruebe la hipótesis de que $\beta_0 = 10$ en el ejercicio 11.8 de la página 399, en comparación con la alternativa de que $\beta_0 < 10$. Utilice un nivel de significancia de 0.05.

11.21 Pruebe la hipótesis de que $\beta_1 = 6$ en el ejercicio 11.9 de la página 399, en comparación con la alternativa de que $\beta_1 < 6$. Utilice un nivel de significancia de 0.025.

11.22 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.16*a* para construir un intervalo de confianza de 95% para $\mu_{Y|85}$ en el ejercicio 11.2 de la página 398.

11.23 Remítase al ejercicio 11.6 de la página 399 y utilice el valor de s^2 que se obtuvo en el ejercicio 11.18*a* para calcular
a) un intervalo de confianza de 95% para la resistencia media al corte cuando $x = 24.5$;
b) un intervalo de predicción de 95% para un solo valor pronosticado de la resistencia al corte cuando $x = 24.5$.

11.24 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.17*a*) y grafique la regresión lineal y las bandas de confianza de 95% para la respuesta media $\mu_{Y|x}$ en el caso de los datos del ejercicio 11.5 de la página 398.

11.25 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.17*a*) y construya un intervalo de confianza de 95% para la cantidad de azúcar convertida correspondiente a $x = 1.6$ en el ejercicio 11.5 de la página 398.

11.26 Remítase al ejercicio 11.3 de la página 398, y utilice el valor de s^2 que se obtuvo en el ejercicio 11.19*a* para calcular
a) un intervalo de confianza de 99% para la cantidad promedio del producto químico que se disolverá en 100 gramos de agua a 50°C;

b) un intervalo de predicción de 99% para la cantidad de producto químico que se disolverá en 100 gramos de agua a 50°C.

11.27 Considere la regresión de la distancia recorrida para ciertos automóviles, en millas por galón (mpg) y su peso en libras (wt). Los datos son de la revista *Consumer Reports* (abril de 1997). En la figura 11.13 se presenta una parte de la salida del *SAS* con los resultados del procedimiento.

a) Estime la distancia recorrida para un vehículo que pesa 4000 libras.

b) Suponga que los ingenieros de Honda afirman que, en promedio, el Civic (o cualquier otro modelo que pese 2440 libras) recorre más de 30 millas por galón (mpg). Con base en los resultados del análisis de regresión, ¿creería usted dicha afirmación? Explique su respuesta.

c) Los ingenieros de diseño del Lexus ES300 consideraron que un rendimiento de 18 mpg sería el objetivo ideal para dicho modelo (o cualquier otro modelo que pese 3390 libras), aunque se espera que haya cierta variación. ¿Es probable que ese objetivo sea realista? Comente al respecto.

11.28 Existen aplicaciones importantes en las que, debido a restricciones científicas conocidas, la recta de regresión **debe atravesar el origen**, es decir, la intersección debe estar en el cero. En otras palabras, el modelo debe ser

$$Y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

y tan sólo se requiere estimar un parámetro sencillo. Con frecuencia a este modelo se le denomina **modelo de regresión por el origen**.

a) Demuestre que el estimador de mínimos cuadrados para la pendiente es

$$b_1 = \left(\sum_{i=1}^{n} x_i y_i\right) \Big/ \left(\sum_{i=1}^{n} x_i^2\right).$$

b) Demuestre que $\sigma^2_{B_1} = \sigma^2 \Big/ \left(\sum_{i=1}^{n} x_i^2\right)$.

c) Demuestre que b_1 del inciso *a* es un estimador insesgado para β_1. Es decir, demuestre que $E(B_1) = \beta_1$.

11.29 Dado el conjunto de datos

y	x
7	2
50	15
100	30
40	10
70	20

a) Grafique los datos.

b) Ajuste una recta de regresión por el origen.

c) Grafique la recta de regresión sobre la gráfica de los datos.

d) Calcule una fórmula general (en términos de y_i y la pendiente b_1) para el estimador de σ^2.

e) Calcule una fórmula para $\text{Var}(\hat{y}_i)$; $i = 1, 2, ..., n$, aplicable a este caso.

f) Grafique límites de confianza de 95% para la respuesta media alrededor de la recta de regresión.

11.30 Para los datos del ejercicio 11.29 calcule un intervalo de predicción de 95% en $x = 25$.

```
                 Root MSE            1.48794   R-Square    0.9509
                 DependentMean      21.50000   Adj R-Sq    0.9447

                              Parameter Estimates
                            Parameter     Standard
          Variable    DF      Estimate        Error   t Value   Pr > |t|
          Intercept    1      44.78018      1.92919     23.21     <.0001
          WT           1      -0.00686   0.00055133    -12.44     <.0001
MODEL       WT    MPG   Predict    LMean     UMean     Lpred     Upred    Residual
GMC       4520     15   13.7720   11.9752   15.5688    9.8988   17.6451    1.22804
Geo       2065     29   30.6138   28.6063   32.6213   26.6385   34.5891   -1.61381
Honda     2440     31   28.0412   26.4143   29.6681   24.2439   31.8386    2.95877
Hyundai   2290     28   29.0703   27.2967   30.8438   25.2078   32.9327   -1.07026
Infiniti  3195     23   22.8618   21.7478   23.9758   19.2543   26.4693    0.13825
Isuzu     3480     21   20.9066   19.8160   21.9972   17.3062   24.5069    0.09341
Jeep      4090     15   16.7219   15.3213   18.1224   13.0158   20.4279   -1.72185
Land      4535     13   13.6691   11.8570   15.4811    9.7888   17.5493   -0.66905
Lexus     3390     22   21.5240   20.4390   22.6091   17.9253   25.1227    0.47599
Lincoln   3930     18   17.8195   16.5379   19.1011   14.1568   21.4822    0.18051
```

Figura 11.13: Salida de resultados del *SAS* para el ejercicio 11.27.

11.7 Selección de un modelo de regresión

Gran parte de lo que se ha presentado hasta ahora acerca de la regresión que involucra una sola variable independiente depende de la suposición de que el modelo elegido es correcto, la suposición de que $\mu_{Y|x}$ se relaciona con x linealmente en los parámetros. Es cierto que no se esperaría que la predicción de la respuesta fuera buena si hubiera diversas variables independientes que no se tomaran en cuenta en el modelo, que afectaran la respuesta y variaran en el sistema. Además, la predicción seguramente sería inadecuada si la estructura verdadera que relaciona $\mu_{Y|x}$ con x fuera extremadamente no lineal en el rango de las variables consideradas.

Es frecuente que se utilice el modelo de regresión lineal simple aun cuando se sepa que el modelo no es lineal o que se desconozca la estructura verdadera. Este método suele ser acertado, en particular cuando el rango de las x es estrecho. De esta manera, el modelo que se utiliza se vuelve una función de aproximación que se espera sea una representación adecuada del panorama verdadero en la región de interés. Sin embargo, hay que señalar el efecto que tendría un modelo inadecuado sobre los resultados presentados hasta este momento. Por ejemplo, si el modelo verdadero, desconocido para el experimentador, es lineal en más de una x, digamos,

$$\mu_{Y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

entonces el estimado $b_1 = S_{xy}/S_{xx}$ de los mínimos cuadrados ordinarios que se calcula considerando tan sólo x_1 en el experimento es, en circunstancias generales, un estimado sesgado del coeficiente β_1, donde el sesgo es una función del coeficiente adicional β_2 (véase el ejercicio de repaso 11.65 en la página 438). Asimismo, el estimado s^2 para σ^2 es sesgado debido a la variable adicional.

11.8 El método del análisis de varianza

Con frecuencia el problema de analizar la calidad de la recta de regresión estimada se maneja por medio del método del **análisis de varianza** (ANOVA), que es un procedimiento mediante el cual la variación total de la variable dependiente se subdivide en componentes significativos, que luego se observan y se tratan en forma sistemática. El análisis de varianza, que se estudia en el capítulo 13, es un recurso poderoso que se emplea en muchas situaciones.

Suponga que tenemos n puntos de datos experimentales en la forma usual (x_i, y_i) y que se estima la recta de regresión. En la sección 11.4 para la estimación de σ^2 se estableció la identidad

$$S_{yy} = b_1 S_{xy} + SCE.$$

Una formulación alternativa y quizá más informativa es la siguiente:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Logramos hacer una partición de la **suma total de los cuadrados corregida de *y*** en dos componentes que deberían proporcionar un significado particular para el experimentador. Esta partición se debería indicar en forma simbólica como

$$STCC = SCR + SCE.$$

El primer componente de la derecha, *SCR*, se denomina **suma de cuadrados de la regresión** y refleja la cantidad de variación de los valores *y* que se **explica con el modelo**, que en este caso es la línea recta postulada. El segundo componente es la ya conocida suma de cuadrados del error, que refleja la variación alrededor de la recta de regresión.

Suponga que nos interesa probar la hipótesis

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } \ H_1\colon\ \beta_1 \neq 0,$$

donde la hipótesis nula en esencia dice que el modelo es $\mu_{Y|x} = \beta_0$; es decir, la variación en los resultados *Y* debida a las fluctuaciones de probabilidad o aleatorias que son independientes de los valores de *x*. Esta condición se refleja en la figura 11.10*b*). En las condiciones de esta hipótesis nula se puede demostrar que SCR/σ^2, y SCE/σ^2 son valores de variables chicuadradas independientes con 1 y $n-2$ grados de libertad, respectivamente y, usando el teorema 7.12, se sigue que $STCC/\sigma^2$ también es un valor de una variable chi cuadrada con $n-1$ grados de libertad. Para probar la hipótesis anterior calculamos

$$f = \frac{SCR\,/\,1}{SCE\,/\,(n-2)} = \frac{SCR}{s^2}$$

y rechazamos H_0 al nivel de significancia α cuando $f > f_\alpha(1, n-2)$.

Por lo general los cálculos se resumen mediante las medias de una **tabla de análisis de varianza**, como se indica en la tabla 11.2. Es costumbre referirse a las distintas sumas de los cuadrados divididos entre sus respectivos grados de libertad como **cuadrados medios**.

Tabla 11.2: Análisis de varianza para la prueba de $\beta_1 = 0$

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**
Regresión	SCR	1	SCR	$\frac{SCR}{s^2}$
Error	SCE	$n-2$	$s^2 = \frac{SCE}{n-2}$	
Total	$STCC$	$n-1$		

Cuando se rechaza la hipótesis nula, es decir, cuando el estadístico *F* calculado excede al valor crítico $f_\alpha(1, n-2)$, concluimos que **hay una cantidad significativa de variación en la respuesta justificada por el modelo postulado, que es la función de la línea recta**. Si el estadístico *F* está en la región de no rechazo, se concluye que los datos no reflejan evidencia suficiente para apoyar el modelo que se postula.

En la sección 11.5 se presentó un procedimiento donde se usa el estadístico

$$T = \frac{B_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$$

para probar la hipótesis

$$H_0\colon\ \beta_1 = \beta_{10} \text{ contra } \ H_1\colon\ \beta_1 \neq \beta_{10},$$

donde *T* sigue la distribución *t* con $n-2$ grados de libertad. La hipótesis se rechaza si $|t| > t_{\alpha/2}$ para un nivel de significancia α. Es interesante observar que en el caso especial en que probamos

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } H_1\colon\ \beta_1 \neq 0,$$

el valor del estadístico T se convierte en

$$t = \frac{b_1}{s/\sqrt{S_{xx}}},$$

y la hipótesis a considerar es idéntica a la que se prueba en la tabla 11.2. En otras palabras, la hipótesis nula establece que la variación en la respuesta se debe tan sólo al azar. El análisis de varianza utiliza la distribución F en vez de la distribución t. Para la alternativa bilateral ambos enfoques son idénticos. Esto se observa si se escribe

$$t^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{b_1 S_{xy}}{s^2} = \frac{SCR}{s^2},$$

que da como resultado un valor idéntico al valor f utilizado en el análisis de varianza. La relación fundamental entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad es

$$t^2 = f(1, v).$$

Desde luego, la prueba t permite probar en comparación con una alternativa unilateral, en tanto que la prueba F está restringida a una prueba en comparación con una alternativa bilateral.

Salida de resultados por computadora comentados para la regresión lineal simple

Considere nuevamente los datos de la tabla 11.1 sobre la reducción de la demanda de oxígeno químico. En las figuras 11.14 y 11.15 se presentan salidas de los resultados por computadora más completos. De nuevo se ilustran con el software *MINITAB*. La columna de la razón t indica pruebas para la hipótesis nula de valores de cero en el parámetro. El término "Fit" denota los valores $\hat{y}$, que con frecuencia se denominan **valores ajustados**. El término "SE Fit" se emplea para calcular los intervalos de confianza sobre la respuesta media. El elemento R^2 se calcula como $(SCR/STCC) \times 100$, y significa la proporción de variación en y explicada por la regresión de la línea recta. Asimismo, se incluyen los intervalos de confianza sobre la respuesta media y los intervalos de predicción sobre una observación nueva.

11.9 Prueba para la linealidad de la regresión: datos con observaciones repetidas

En ciertos tipos de situaciones experimentales el investigador tiene la capacidad de efectuar observaciones repetidas de la respuesta para cada valor de x. Aunque no es necesario tener dichas repeticiones para estimar β_0 y β_1, las repeticiones permiten al experimentador obtener información cuantitativa acerca de lo apropiado que resulta el modelo. De hecho, si se generan observaciones repetidas, el investigador puede efectuar una prueba de significancia para determinar si el modelo es o no adecuado.

```
The regression equation is COD = 3.83 + 0.904 Per_Red
Predictor      Coef   SE Coef       T       P
 Constant    3.830     1.768    2.17   0.038
  Per_Red  0.90364   0.05012   18.03   0.000
S = 3.22954    R-Sq = 91.3%    R-Sq(adj ) = 91.0%
              Analysis of Variance
Source            DF       SS       MS        F       P
Regression         1   3390.6   3390.6   325.08   0.000
Residual  Error 31    323.3     10.4
Total             32   3713.9

Obs   Per_Red      COD      Fit   SE Fit   Residual   St Resid
  1       3.0    5.000    6.541    1.627     -1.541      -0.55
  2      36.0   34.000   36.361    0.576     -2.361      -0.74
  3       7.0   11.000   10.155    1.440      0.845       0.29
  4      37.0   36.000   37.264    0.590     -1.264      -0.40
  5      11.0   21.000   13.770    1.258      7.230       2.43
  6      38.0   38.000   38.168    0.607     -0.168      -0.05
  7      15.0   16.000   17.384    1.082     -1.384      -0.45
  8      39.0   37.000   39.072    0.627     -2.072      -0.65
  9      18.0   16.000   20.095    0.957     -4.095      -1.33
 10      39.0   36.000   39.072    0.627     -3.072      -0.97
 11      27.0   28.000   28.228    0.649     -0.228      -0.07
 12      39.0   45.000   39.072    0.627      5.928       1.87
 13      29.0   27.000   30.035    0.605     -3.035      -0.96
 14      40.0   39.000   39.975    0.651     -0.975      -0.31
 15      30.0   25.000   30.939    0.588     -5.939      -1.87
 16      41.0   41.000   40.879    0.678      0.121       0.04
 17      30.0   35.000   30.939    0.588      4.061       1.28
 18      42.0   40.000   41.783    0.707     -1.783      -0.57
 19      31.0   30.000   31.843    0.575     -1.843      -0.58
 20      42.0   44.000   41.783    0.707      2.217       0.70
 21      31.0   40.000   31.843    0.575      8.157       2.57
 22      43.0   37.000   42.686    0.738     -5.686      -1.81
 23      32.0   32.000   32.746    0.567     -0.746      -0.23
 24      44.0   44.000   43.590    0.772      0.410       0.13
 25      33.0   34.000   33.650    0.563      0.350       0.11
 26      45.0   46.000   44.494    0.807      1.506       0.48
 27      33.0   32.000   33.650    0.563     -1.650      -0.52
 28      46.0   46.000   45.397    0.843      0.603       0.19
 29      34.0   34.000   34.554    0.563     -0.554      -0.17
 30      47.0   49.000   46.301    0.881      2.699       0.87
 31      36.0   37.000   36.361    0.576      0.639       0.20
 32      50.0   51.000   49.012    1.002      1.988       0.65
 33      36.0   38.000   36.361    0.576      1.639       0.52
```

Figura 11.14: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte I.

Seleccionemos una muestra aleatoria de n observaciones utilizando k valores distintos de x, por ejemplo, $x_1, x_2, ..., x_n$, tales que la muestra contenga n_1 valores observados de la variable aleatoria Y_1 correspondientes a los valores x_1, con n_2 valores observados de Y_2 correspondientes a $x_2, ..., n_k$ valores observados de Y_k correspondientes a x_k. Necesariamente, $n = \sum_{i=1}^{k} n_i$.

Obs	Fit	SE Fit	95% CI	95% PI
1	6.541	1.627	(3.223, 9.858)	(-0.834, 13.916)
2	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
3	10.155	1.440	(7.218, 13.092)	(2.943, 17.367)
4	37.264	0.590	(36.062, 38.467)	(30.569, 43.960)
5	13.770	1.258	(11.204, 16.335)	(6.701, 20.838)
6	38.168	0.607	(36.931, 39.405)	(31.466, 44.870)
7	17.384	1.082	(15.177, 19.592)	(10.438, 24.331)
8	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
9	20.095	0.957	(18.143, 22.047)	(13.225, 26.965)
10	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
11	28.228	0.649	(26.905, 29.551)	(21.510, 34.946)
12	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
13	30.035	0.605	(28.802, 31.269)	(23.334, 36.737)
14	39.975	0.651	(38.648, 41.303)	(33.256, 46.694)
15	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
16	40.879	0.678	(39.497, 42.261)	(34.149, 47.609)
17	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
18	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
19	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
20	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
21	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
22	42.686	0.738	(41.181, 44.192)	(35.930, 49.443)
23	32.746	0.567	(31.590, 33.902)	(26.059, 39.434)
24	43.590	0.772	(42.016, 45.164)	(36.818, 50.362)
25	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
26	44.494	0.807	(42.848, 46.139)	(37.704, 51.283)
27	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
28	45.397	0.843	(43.677, 47.117)	(38.590, 52.205)
29	34.554	0.563	(33.406, 35.701)	(27.868, 41.239)
30	46.301	0.881	(44.503, 48.099)	(39.473, 53.128)
31	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
32	49.012	1.002	(46.969, 51.055)	(42.115, 55.908)
33	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)

Figura 11.15: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte II.

Definimos

$$y_{ij} = \text{el } j\text{-ésimo valor de la variable aleatoria } Y_i,$$
$$y_{i.} = T_{i.} = \sum_{j=1}^{n_i} y_{ij},$$
$$\bar{y}_{i.} = \frac{T_{i.}}{n_i}.$$

Entonces, si se realizaron $n_4 = 3$ mediciones de Y que corresponden a $x = x_4$, estas observaciones se indicarían por medio de y_{41}, y_{42} y y_{43}. Por lo tanto,

$$T_{i.} = y_{41} + y_{42} + y_{43}.$$

El concepto de la falta de ajuste

La suma de cuadrados del error consta de dos partes: la cantidad debida a la variación entre los valores de Y dentro de valores dados de x, y un componente que normalmente

se denomina contribución a la **falta de ajuste**. El primer componente refleja tan sólo la variación aleatoria, o **error experimental puro**, en tanto que el segundo es una medida de la variación sistemática introducida por los términos de orden superior. En nuestro caso éstos son términos de x distintos de la contribución lineal o de primer orden. Observe que al elegir un modelo lineal en esencia asumimos que este segundo componente no existe y que, en consecuencia, la suma de cuadrados del error se debe por completo a errores aleatorios. Si éste fuera el caso, entonces $s^2 = SCE/(n-2)$ es un estimado insesgado de σ^2. Sin embargo, si el modelo no se ajusta a los datos en forma apropiada, entonces la suma de cuadrados del error estará inflada y producirá un estimador sesgado de σ^2. Ya sea que el modelo se ajuste o no a los datos, siempre que se tienen observaciones repetidas es posible obtener un estimador insesgado de σ^2 calculando

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}, \quad i = 1, 2, \ldots, k,$$

para cada uno de los k valores distintos de x y, después, agrupando estas varianzas, tenemos

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n-k}.$$

El numerador de s^2 es una **medida del error experimental puro**. A continuación se presenta un procedimiento de cálculo para separar la suma de los cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste:

Cálculo de la suma de los cuadrados de la falta de ajuste

1. Calcular la suma de los cuadrados del error puro

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

Esta suma de cuadrados tiene $n-k$ grados de libertad asociados con ella, y el cuadrado medio resultante es el estimador insesgado s^2 de σ^2.

2. Restar la suma de los cuadrados del error puro de la suma de los cuadrados del error, *SCE*, con lo que se obtiene la suma de los cuadrados debida a la falta de ajuste. Los grados de libertad de la falta de ajuste también se obtienen simplemente restando $(n-2) - (n-k) = k-2$.

Los cálculos necesarios para probar hipótesis en un problema de regresión con mediciones repetidas de la respuesta se pueden resumir como se muestra en la tabla 11.3.

Las figuras 11.16 y 11.17 ilustran los puntos muestrales para las situaciones del "modelo correcto" y del "modelo incorrecto". En la figura 11.16, donde $\mu_{Y|x}$ cae sobre una línea recta, no hay falta de ajuste cuando se asume un modelo lineal, por lo que la variación muestral alrededor de la recta de regresión es un error puro que resulta de la variación que ocurre entre observaciones repetidas. En la figura 11.17, donde es evidente que $\mu_{Y|x}$ no cae sobre una línea recta, la responsable de la mayor parte de la variación alrededor de la recta de regresión, además del error puro, es la falta de ajuste que resulta de seleccionar por error un modelo lineal.

Tabla 11.3: Análisis de varianza para la prueba de linealidad de la regresión

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
Regresión	SCR	1	SCR	$\frac{SCR}{s^2}$
Error	SCE	$n-2$		
Falta de ajuste	SCE-SCE (puro)	$k-2$	$\frac{SCE\text{-}SCE\ (\text{puro})}{k-2}$	$\frac{SCE\text{-}SCE\ (\text{puro})}{s^2(k-2)}$
Error puro	SCE (puro)	$n-k$	$s^2 = \frac{SCE\ (\text{puro})}{n-k}$	
Total	$STCC$	$n-1$		

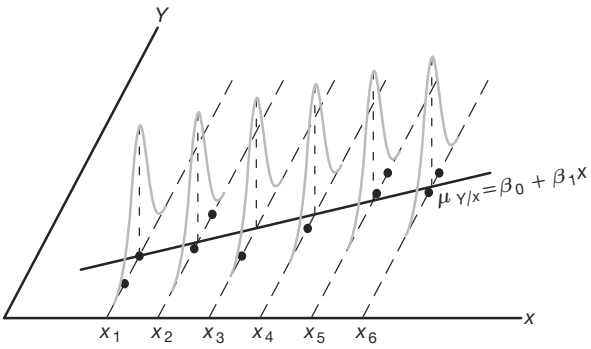


Figura 11.16: Modelo lineal correcto con componente sin falta de ajuste.

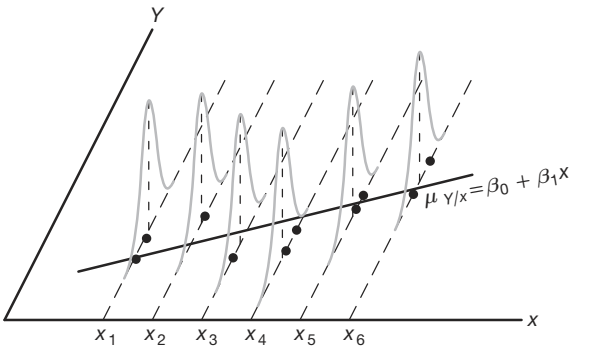


Figura 11.17: Modelo lineal incorrecto con componente de falta de ajuste.

¿Por qué es importante detectar la falta de ajuste?

El concepto de falta de ajuste es muy importante en las aplicaciones del análisis de regresión. De hecho, la necesidad de construir o diseñar un experimento que tome en cuenta la falta de ajuste se vuelve más crítica a medida que el problema y el mecanismo subyacente implicados se vuelven más complicados. Es cierto que no siempre se puede tener la certeza de que la estructura que se postula, en este caso el modelo de regresión lineal, sea una representación correcta o incluso adecuada. El ejemplo siguiente muestra la manera en que se parte la suma de cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste. Lo adecuado del modelo se prueba al nivel de significancia α, comparando el cuadrado medio de la falta de ajuste dividido entre s^2 con $f_\alpha(k-2, n-k)$.

Ejemplo 11.8: En la tabla 11.4 se presenta el registro de las observaciones del producto de una reacción química tomadas a distintas temperaturas. Calcule el modelo lineal $\mu_{Y|x} = \beta_0 + \beta_1 x$ y pruebe la falta de ajuste.

***Solución*:** Los resultados de los cálculos se presentan en la tabla 11.5.

Conclusión: La partición de la variación total de esta manera revela una variación significativa debida al modelo lineal y una cantidad insignificante de variación debida a la falta de ajuste. Por consiguiente, los datos experimentales no parecen sugerir la necesidad de considerar en el modelo términos superiores a los de primer orden y no se rechaza la hipótesis nula. ∎

Tabla 11.4: Datos para el ejemplo 11.8

y (%)	x (°C)	y (%)	x (°C)
77.4	150	88.9	250
76.7	150	89.2	250
78.2	150	89.7	250
84.1	200	94.8	300
84.5	200	94.7	300
83.7	200	95.9	300

Tabla 11.5: Análisis de varianza de los datos de producto-temperatura

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada	Valores P
Regresión	509.2507	1	509.2507	1531.58	< 0.0001
Error	3.8660	10			
Falta de ajuste	1.2060	2	0.6030	1.81	0.2241
Error puro	2.6600	8	0.3325		
Total	513.1167	11			

Salida de resultados por computadora comentados para la prueba de falta de ajuste

En la figura 11.18 se presenta una salida de resultados por computadora para el análisis de los datos del ejemplo 11.8 con el programa *SAS*. Observe la "LOF" con 2 grados de libertad, que representa las contribuciones cuadrática y cúbica al modelo, y el valor P de 0.22, que sugiere que el modelo lineal (de primer orden) es adecuado.

```
Dependent Variable: yield
                                      Sum of
 Source                    DF         Squares     Mean Square    F Value    Pr > F
 Model                      3     510.4566667     170.1522222     511.74    <.0001
 Error                      8       2.6600000       0.3325000
 Corrected Total           11     513.1166667
              R-Square     Coeff Var      Root MSE    yield Mean
              0.994816      0.666751      0.576628      86.48333
 Source                    DF       Type I SS     Mean Square    F Value    Pr > F
 temperature                1     509.2506667     509.2506667    1531.58    <.0001
 LOF                        2       1.2060000       0.6030000       1.81    0.2241
```

Figura 11.18: Salida de resultados del *SAS* que incluye el análisis de los datos del ejemplo 11.8.

Ejercicios

11.31 En el ejercicio 11.3 de la página 398 pruebe la linealidad de la regresión. Use un nivel de significancia de 0.05. Haga comentarios al respecto.

11.32 En el ejercicio 11.8 de la página 399 pruebe la linealidad de la regresión. Haga comentarios al respecto.

11.33 Suponga que tenemos una ecuación lineal que pasa por el origen $\mu_{Y|x} = \beta x$ (ejercicio 11.28).

a) Estime la regresión lineal que pasa por el origen para los siguientes datos:

x	0.5	1.5	3.2	4.2	5.1	6.5
y	1.3	3.4	6.7	8.0	10.0	13.2

b) Suponga que se desconoce si la regresión verdadera debería pasar por el origen. Estime el modelo lineal $\mu_{Y|x} = \beta_0 + \beta_1 x$ y pruebe la hipótesis de que $\beta_0 = 0$ a un nivel de significancia de 0.10, en comparación con la alternativa de que $\beta_0 \neq 0$.

11.34 En el ejercicio 11.5 de la página 398 utilice el método del análisis de varianza para probar la hipótesis de que $\beta_1 = 0$, en comparación con la hipótesis alternativa de que $\beta_1 \neq 0$, a un nivel de significancia de 0.05.

11.35 Los siguientes datos son el resultado de una investigación sobre el efecto de la temperatura de reacción x sobre la conversión porcentual de un proceso químico y. (Véase Myers, Montgomery y Anderson-Cook, 2009). Ajuste una regresión lineal simple y utilice pruebas de falta de ajuste para determinar si el modelo es adecuado. Analice los resultados.

Observación	Temperatura (°C), x	Conversión (%), y
1	200	43
2	250	78
3	200	69
4	250	73
5	189.65	48
6	260.35	78
7	225	65
8	225	74
9	225	76
10	225	79
11	225	83
12	225	81

11.36 La ganancia de un transistor en un dispositivo de circuito integrado, entre el emisor y el colector (hFE), se relaciona con dos variables (Myers, Montgomery y Anderson-Cook, 2009) que se controlan en el proceso de deposición, controlado por el emisor en el tiempo (x_1, en minutos) y la dosis del emisor (x_2, en iones $\times 10^{14}$). Se observaron 14 muestras después de la deposición y los datos resultantes se presentan en la tabla siguiente. Consideraremos modelos de regresión lineal usando la ganancia como respuesta y el control del emisor en el tiempo o la dosis del emisor como la variable regresora.

Obs.	x_1 (tiempo de control, min)	x_2 (dosis, iones $\times 10^{14}$)	y (ganancia o hFE)
1	195	4.00	1004
2	255	4.00	1636
3	195	4.60	852
4	255	4.60	1506
5	255	4.20	1272
6	255	4.10	1270
7	255	4.60	1269
8	195	4.30	903
9	255	4.30	1555
10	255	4.00	1260
11	255	4.70	1146
12	255	4.30	1276
13	255	4.72	1225
14	340	4.30	1321

a) Determine si el tiempo de control del emisor influye en la ganancia en una relación lineal. Es decir, pruebe H_0: $\beta_1 = 0$, donde β_1 es la pendiente de la variable regresora.

b) Efectúe una prueba de falta de ajuste para determinar si la relación lineal es adecuada. Saque sus conclusiones.

c) Determine si la dosis del emisor influye en la ganancia en una relación lineal. ¿Cuál variable regresora es el mejor predictor de la ganancia?

11.37 En los pesticidas se utilizan compuestos de organofosfatos (OF). Sin embargo, es importante estudiar el efecto que tienen sobre las especies expuestas a ellos. Como parte del estudio de laboratorio *Some Effects of Organophosphate Pesticides on Wildlife Species*, elaborado por el Departamento de Pesca y Vida Silvestre de Virginia Tech, se realizó un experimento en el cual se suministraron distintas dosis de un pesticida de OF específico a 5 grupos de 5 ratones (*peromysius leucopus*). Los 25 ratones eran hembras de edad y condiciones similares. Un grupo no recibió el producto. La respuesta básica y consistió en medir la actividad cerebral. Se postuló que dicha actividad disminuiría con un incremento en la dosis de OF. A continuación se presentan los datos:

Animal	Dosis, x (mg/kg de peso corporal)	Actividad, y (moles/litro/min)
1	0.0	10.9
2	0.0	10.6
3	0.0	10.8
4	0.0	9.8
5	0.0	9.0
6	2.3	11.0
7	2.3	11.3
8	2.3	9.9
9	2.3	9.2
10	2.3	10.1
11	4.6	10.6
12	4.6	10.4
13	4.6	8.8
14	4.6	11.1
15	4.6	8.4
16	9.2	9.7
17	9.2	7.8
18	9.2	9.0
19	9.2	8.2
20	9.2	2.3
21	18.4	2.9
22	18.4	2.2
23	18.4	3.4
24	18.4	5.4
25	18.4	8.2

a) Con el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2,...,25,$$

calcule los estimados de los mínimos cuadrados de β_0 y β_1.

b) Construya una tabla de análisis de varianza en la cual aparezcan por separado el error puro y el error por falta de ajuste. Determine si la falta de ajuste es significativa al nivel de 0.05. Interprete los resultados.

11.38 Es frecuente que se utilice el tratamiento con calor para carburar partes metálicas como los engranes. El espesor de la capa carburada se considera una característica importante del engrane que contribuye a la confiabilidad general de la parte. Debido a la naturaleza crítica de esta característica, se realiza una prueba de laboratorio para cada lote del horno. La prueba es destructiva, ya que una parte real se corta en forma transversal y se sumerge en un producto químico durante cierto tiempo. Esta prueba requiere que se efectúe un análisis del carbono sobre la superficie, tanto de la parte superior del engrane (arriba de los dientes) como de su raíz (entre los dientes). Los datos siguientes son los resultados de la prueba de análisis de carbono en 19 partes.

Tiempo de inmersión	Grado	Tiempo de inmersión	Grado
0.58	0.013	1.17	0.021
0.66	0.016	1.17	0.019
0.66	0.015	1.17	0.021
0.66	0.016	1.20	0.025
0.66	0.015	2.00	0.025
0.66	0.016	2.00	0.026
1.00	0.014	2.20	0.024
1.17	0.021	2.20	0.025
1.17	0.018	2.20	0.024
1.17	0.019		

a) Ajuste una regresión lineal simple que relacione el grado del análisis de carbono *y* en comparación con el tiempo de inmersión. Pruebe H_0: $\beta_1 = 0$.

b) Si se rechaza la hipótesis del inciso *a*, determine si el modelo lineal es adecuado.

11.39 Se desea obtener un modelo de regresión que relacione la temperatura con la proporción de impurezas de una sustancia que pasa a través de helio sólido. Se lista la temperatura en grados centígrados. A continuación se presentan los datos.

Temperatura (°C)	Proporción de impurezas
−260.5	0.425
−255.7	0.224
−264.6	0.453
−265.0	0.475
−270.0	0.705
−272.0	0.860
−272.5	0.935
−272.6	0.961
−272.8	0.979
−272.9	0.990

a) Ajuste un modelo de regresión lineal.

b) ¿Parece que la proporción de impurezas que pasan a través del helio aumenta a medida que la temperatura se acerca a −273 grados centígrados?

c) Calcule R^2.

d) Con base en la información anterior, ¿parece adecuado el modelo lineal? ¿Qué información adicional necesitaría usted para responder mejor a la pregunta?

11.40 Existe interés por estudiar el efecto que tiene el tamaño de la población de varias ciudades de Estados Unidos sobre las concentraciones de ozono. Los datos consisten en la población de 1999 en millones de habitantes y en la cantidad de ozono presente por hora en partes por mil millones (ppmm). Los datos son los siguientes:

Ozono (ppmm/hora), *y*	Población, *x*
126	0.6
135	4.9
124	0.2
128	0.5
130	1.1
128	0.1
126	1.1
128	2.3
128	0.6
129	2.3

a) Ajuste un modelo de regresión lineal que relacione la concentración de ozono con la población. Pruebe H_0: $\beta_1 = 0$ usando el método ANOVA.

b) Haga una prueba para la falta de ajuste. Con base en los resultados de la prueba, ¿es apropiado el modelo lineal?

c) Pruebe la hipótesis del inciso *a*) utilizando el cuadrado medio del error puro en la prueba *F*. ¿Cambian los resultados? Comente las ventajas de cada prueba.

11.41 Evaluar la deposición del nitrógeno de la atmósfera es una tarea importante del National Atmospheric Deposition Program (NADP), que está asociado con muchas instituciones. Este programa está estudiando la deposición atmosférica y su efecto sobre los cultivos agrícolas, las aguas superficiales de los bosques y otros recursos. Los óxidos del nitrógeno pueden tener efectos sobre el ozono atmosférico y la cantidad de nitrógeno puro que se encuentra en el aire que respiramos. Los datos son los siguientes:

Año	Óxido de nitrógeno
1978	0.73
1979	2.55
1980	2.90
1981	3.83
1982	2.53
1983	2.77
1984	3.93
1985	2.03
1986	4.39
1987	3.04
1988	3.41
1989	5.07
1990	3.95
1991	3.14
1992	3.44
1993	3.63
1994	4.50
1995	3.95
1996	5.24
1997	3.30
1998	4.36
1999	3.33

a) Grafique los datos.

b) Ajuste un modelo de regresión lineal y calcule R^2.

c) ¿Qué puede decir acerca de la tendencia del óxido de nitrógeno con el paso del tiempo?

11.42 Para una variedad particular de planta los investigadores desean desarrollar una fórmula para predecir la cantidad de semillas (en gramos) como una función de la densidad de las plantas. Efectuaron un estudio con cuatro niveles del factor x, el número de plantas por parcela. Se utilizaron cuatro réplicas para cada nivel de x. A continuación se muestran los datos:

Plantas por parcela, x	Cantidad de semillas, y (gramos)			
10	12.6	11.0	12.1	10.9
20	15.3	16.1	14.9	15.6
30	17.9	18.3	18.6	17.8
40	19.2	19.6	18.9	20.0

¿Es adecuado un modelo de regresión lineal simple para analizar este conjunto de datos?

11.10 Gráficas de datos y transformaciones

En este capítulo se estudia la construcción de modelos de regresión en los que hay una variable independiente o regresora. Además, se supone que durante la construcción del modelo tanto x como y entran en el modelo en *forma lineal*. Con frecuencia es aconsejable trabajar con un modelo alternativo en el que x o y (o ambas) intervengan en una forma no lineal. Se podría recomendar una **transformación** de los datos debido a consideraciones teóricas inherentes al estudio científico, o bien, una simple graficación de los datos podría sugerir la necesidad de *reexpresar* las variables en el modelo. La necesidad de llevar a cabo una transformación es muy fácil de diagnosticar en el caso de la regresión lineal simple, ya que las gráficas en dos dimensiones brindan un panorama verdadero de la manera en que las variables se comportan en el modelo.

Un modelo en el que x o y se transforman no debería considerarse como un *modelo de regresión no lineal*. Por lo general denominamos a un modelo de regresión como lineal cuando es **lineal en los parámetros**. En otras palabras, suponga que el aspecto de los datos u otra información científica sugiere que debe hacerse la **regresión de y^* en comparación con la de x^***, donde cada una de ellas es una transformación de las variables naturales x y y. Entonces, el modelo de la forma

$$y_i^* = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

es lineal porque lo es en los parámetros β_0 y β_1. El material que se estudió en las secciones 11.2 a 11.9 permanece sin cambio, donde y_i^* y x_i^* reemplazan a y_i y x_i. Un ejemplo sencillo y útil es el modelo log-log:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i.$$

Aunque este modelo es no lineal en x y y, sí lo es en los parámetros y por ello recibe el tratamiento de un modelo lineal. Por otro lado, un ejemplo de modelo verdaderamente no lineal es:

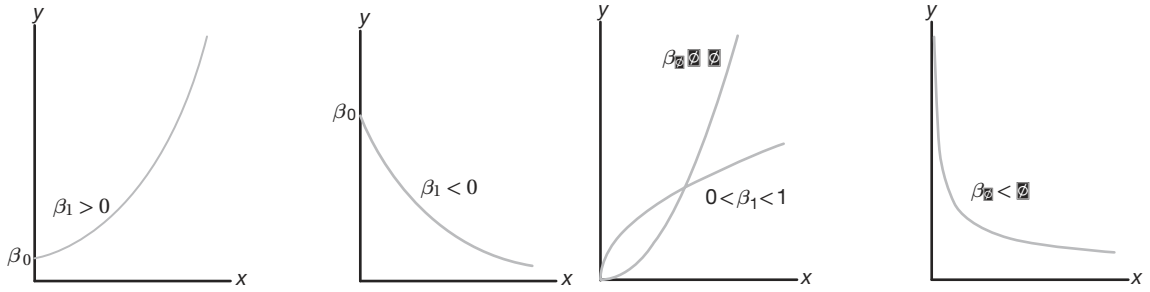
$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \epsilon_i,$$

donde se debe estimar el parámetro β_2, así como β_0 y β_1. El modelo es no lineal en β_2.

Las transformaciones susceptibles de mejorar el ajuste y la capacidad de predicción de un modelo son muy numerosas. Para un análisis completo de las transformaciones el lector podría consultar a Myers (1990, véase la bibliografía). Decidimos incluir aquí algunas de ellas y mostrar la apariencia de las gráficas que sirven como herramientas diagnósticas. Considere la tabla 11.6, donde se presentan varias funciones que describen relaciones entre y y x que pueden producir una *regresión lineal* por medio de la transformación indicada. Además, en aras de que el análisis sea más exhaustivo, se presentan al lector las variables dependiente e independiente que se utilizan en la *regresión lineal simple* resultante. La figura 11.19 ilustra las funciones que se listan en la tabla 11.6, las cuales sirven como guía para el analista en la elección de una transformación a partir de la observación de la gráfica de y contra x.

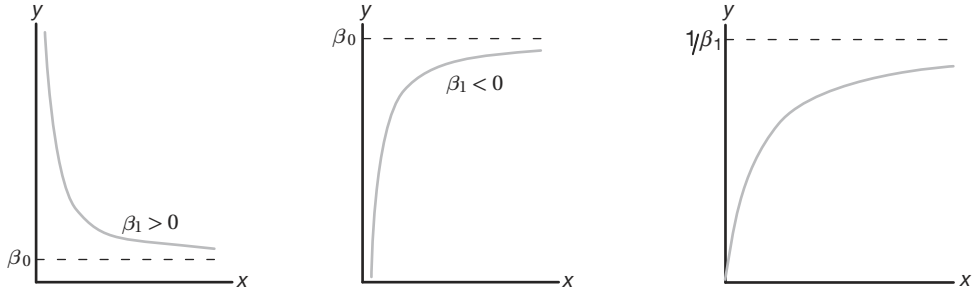
Tabla 11.6: Algunas transformaciones útiles para linealizar

Forma funcional que relaciona y con x	**Transformación propia**	**Forma de la regresión lineal simple**
Exponencial: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Hacer la regresión de y^* contra x
Potencia: $y = \beta_0 x^{\beta_1}$	$y^* = \log y$; $x^* = \log x$	Hacer la regresión de y^* contra x^*
Recíproca: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Hacer la regresión de y contra x^*
Hiperbólica: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}$; $x^* = \frac{1}{x}$	Hacer la regresión de y^* contra x^*



a) Función exponencial

b) Función de potencia



c) Función recíproca

d) Función hiperbólica

Figura 11.19: Diagramas que ilustran las funciones listadas en la tabla 11.6.

¿Cuáles son las implicaciones de un modelo transformado?

Lo que sigue intenta ser una ayuda para el analista cuando es evidente que una transformación producirá una mejoría. Sin embargo, antes de dar un ejemplo hay que mencionar dos puntos importantes. El primero tiene que ver con la escritura formal del modelo una vez que se hayan transformado los datos. Con mucha frecuencia el analista no piensa en esto y simplemente lleva a cabo la transformación sin preocuparse por la forma del modelo *antes* ni *después* de la transformación. El modelo exponencial sirve como una buena ilustración de esto. El modelo en las variables naturales (no transformadas) que produce un *modelo de error aditivo* en las variables transformadas es dado por

$$y_i = \beta_0 e^{\beta_1 x_i} \cdot \epsilon_i,$$

que es un *modelo de error multiplicativo*. Al aplicar logaritmos es claro que se obtiene

$$\ln y_i = \ln \beta_0 + \beta_1 x_i + \ln \epsilon_i.$$

Como resultado, las suposiciones básicas se efectúan sobre $\ln \epsilon_i$. El propósito de esta presentación sólo es recordar al lector que no debemos considerar una transformación tan sólo como una manipulación algebraica a la cual se suma un error. Con frecuencia, un modelo en las variables transformadas que tiene una adecuada *estructura de error aditivo* es resultado de un modelo en las variables naturales con un tipo de estructura de error diferente.

El segundo aspecto importante se refiere a la noción de las medidas de mejoría. Las medidas evidentes de comparación son, por supuesto, el valor de R^2 y el cuadrado medio de los residuales s^2. (En el capítulo 12 se estudian otras medidas de rendimiento que se usan para comparar modelos que compiten). Ahora, si la respuesta y no se transforma, entonces es claro que s^2 y R^2 se pueden usar para medir la utilidad de la transformación. Los residuales estarán en las mismas unidades para los dos modelos, el transformado y el que no se transformó. No obstante, cuando se transforma y los criterios de rendimiento para el modelo transformado deberían basarse en los valores de los residuales en las unidades de medida de la respuesta no transformada. De esta manera las comparaciones son más apropiadas. El siguiente ejemplo proporciona una ilustración de lo anterior.

Ejemplo 11.9: Se registra la presión P de un gas que corresponde a distintos volúmenes V y los datos se presentan en la tabla 11.7.

Tabla 11.7: Datos para el ejemplo 11.9

V (cm^3)	50	60	70	90	100
P (kg/cm^2)	64.7	51.3	40.5	25.9	7.8

La ley del gas ideal es dada por la forma funcional $PV^\gamma = C$, donde γ y C son constantes. Estime las constantes C y γ.

***Solución*:** Se toman logaritmos naturales en ambos lados del modelo

$$P_i V^\gamma = C \cdot \epsilon_i, \quad i = 1, 2, 3, 4, 5.$$

Como resultado, es posible escribir el modelo lineal

$$\ln P_i = \ln C - \gamma \ln V_i + \epsilon_i^*, \quad \text{i} = 1, 2, 3, 4, 5,$$

Donde $\epsilon_i^* = \ln \epsilon_i$. Los siguientes son los resultados de la regresión lineal simple:

Intersección $\widehat{\ln C} = 14.7589$, $\widehat{C} = 2{,}568{,}862.88$, Pendiente: $\hat{\gamma} = 2.65347221$.

La siguiente tabla representa información tomada del análisis de regresión.

P_i	V_i	$\ln P_i$	$\ln V_i$	$\widehat{\ln P_i}$	$\widehat{P_i}$	$e_i = P_i - \widehat{P_i}$
64.7	50	4.16976	3.91202	4.37853	79.7	−15.0
51.3	60	3.93769	4.09434	3.89474	49.1	2.2
40.5	70	3.70130	4.24850	3.48571	32.6	7.9
25.9	90	3.25424	4.49981	2.81885	16.8	9.1
7.8	100	2.05412	4.60517	2.53921	12.7	−4.9

Resulta aleccionador graficar los datos y la ecuación de regresión. En la figura 11.20 se presenta una gráfica de los datos no transformados de presión y volumen; en tanto que la curva representa la ecuación de regresión.

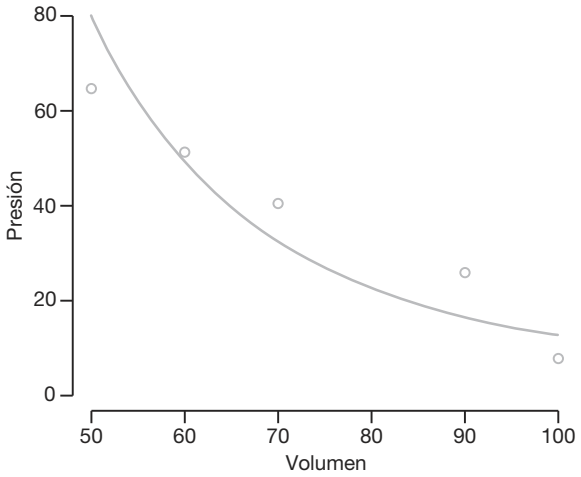


Figura 11.20: Datos de presión y volumen y la regresión ajustada.

Gráficas de diagnóstico de los residuales: detección gráfica de la transgresión de las suposiciones

Las gráficas de los datos brutos pueden ser muy útiles para determinar la naturaleza del modelo que debe ajustarse a ellos cuando sólo hay una variable independiente. En lo anterior tratamos de ilustrar esto. Sin embargo, la detección de la forma del modelo adecuado no es el único beneficio que se obtiene con la gráfica de diagnóstico. Como ocurre con gran parte del material asociado con las pruebas de hipótesis que se expone en el capítulo 10, los métodos de graficación ilustran y detectan la transgresión de las suposiciones. El lector debería recordar que muchos de los conceptos que se ilustran en este capítulo requieren suposiciones sobre los errores del modelo, las ϵ_i. De hecho, suponemos que las ϵ_i son variables aleatorias independientes $N(0, \sigma)$. Por supuesto, las ϵ_i no se observan. Sin embargo, las $e_i = y_i - \hat{y}_i$, los *residuales*, corresponden al error en el ajuste de la recta de regresión, por lo que sirven para imitar a las ϵ_i. Así, la apariencia general de estos residuales con frecuencia puede resaltar las dificultades. De manera ideal, por supuesto, la gráfica de los residuales es como la que se aprecia en la figura 11.21. Es decir, los residuales deberían demostrar en verdad fluctuaciones aleatorias alrededor del valor de cero.

Varianza no homogénea

Una suposición importante que se hace en el análisis de regresión es la varianza homogénea. A menudo las transgresiones se detectan mediante la apariencia de la gráfica de residuales. Es común que en los datos científicos se incremente la varianza del error con el aumento de la variable regresora. Una varianza grande del error produce residuales grandes y, por ende, una gráfica de residuales como la que se presenta en la figura 11.22 es una señal de varianza no homogénea. En el capítulo 12, en el cual se expone la regresión lineal múltiple, se presenta un análisis más amplio acerca de las gráficas de los residuales e información acerca de los diferentes tipos de residuales.

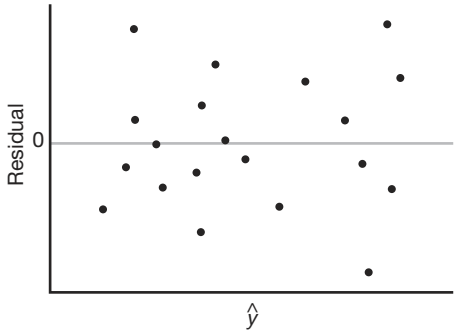


Figura 11.21: Gráfica ideal de los residuales.

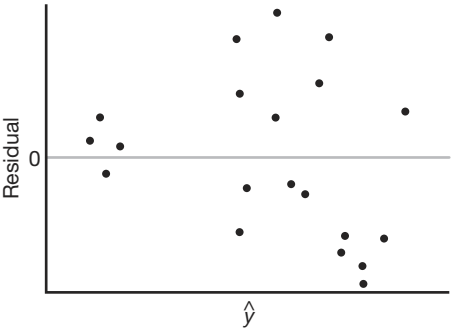


Figura 11.22: Gráfica de los residuales que ilustra una varianza heterogénea del error.

Gráfica de la probabilidad normal

La suposición de que los errores del modelo son normales se hace cuando el analista de los datos se ocupa de las pruebas de hipótesis o de la estimación de intervalos de confianza. De nuevo, los equivalentes numéricos de los ϵ_i, es decir, los residuales, son sujetos de diagnóstico mediante la graficación para detectar cualesquiera transgresiones extremas. En el capítulo 8 se presentaron las gráficas normales cuantil-cuantil y se analizaron en forma breve las de probabilidad normal. En el estudio de caso que se presenta en la siguiente sección se ilustran estas gráficas de residuales.

11.11 Estudio de caso de regresión lineal simple

En la fabricación de productos comerciales de madera es importante estimar la relación que hay entre la densidad de un producto de madera y su rigidez. Se está considerando un tipo relativamente nuevo de aglomerado que se puede formar con mucha mayor facilidad que el producto comercial ya aceptado. Es necesario saber a qué densidad su rigidez es comparable con la del producto comercial bien conocido y documentado. Terrance E. Conners realizó un estudio titulado *Investigation of Certain Mechanical Properties of a Wood-Foam Composite* (Tesis para el doctorado, Departamento de Bosques y Vida Silvestre, University of Massachusetts). Se produjeron 30 tableros de aglomerado con densidades que variaban aproximadamente de 8 a 26 libras por pie cúbico y se midió su rigidez en libras por pulgada cuadrada. En la tabla 11.8 se presentan los datos.

Es necesario que el analista de datos se concentre en un ajuste apropiado para los datos y que utilice los métodos de inferencia que se estudian en este capítulo. Tal vez lo más apropiado sea una prueba de hipótesis sobre la pendiente de la regresión, así como

la estimación de los intervalos de confianza o de predicción. Se comenzará presentando un simple diagrama de dispersión de los datos brutos con una regresión lineal simple sobrepuesta. En la figura 11.23 se observa dicha gráfica.

El ajuste de regresión lineal simple a los datos produce el modelo ajustado

$$\hat{y} = -25{,}433.739 + 3884.976x \quad (R^2 = 0.7975),$$

Tabla 11.8: Densidad y rigidez de 30 tableros de aglomerado

Densidad, *x*	**Rigidez, *y***	**Densidad, *x***	**Rigidez, *y***
9.50	14,814.00	8.40	17,502.00
9.80	14,007.00	11.00	19,443.00
8.30	7573.00	9.90	14,191.00
8.60	9714.00	6.40	8076.00
7.00	5304.00	8.20	10,728.00
17.40	43,243.00	15.00	25,319.00
15.20	28,028.00	16.40	41,792.00
16.70	49,499.00	15.40	25,312.00
15.00	26,222.00	14.50	22,148.00
14.80	26,751.00	13.60	18,036.00
25.60	96,305.00	23.40	104,170.00
24.40	72,594.00	23.30	49,512.00
19.50	32,207.00	21.20	48,218.00
22.80	70,453.00	21.70	47,661.00
19.80	38,138.00	21.30	53,045.00

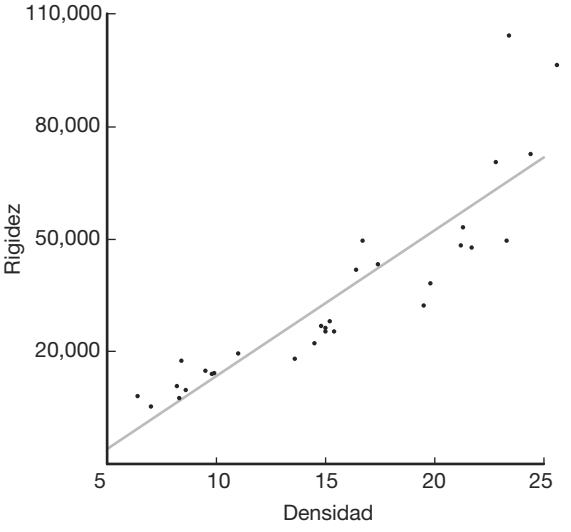


Figura 11.23: Diagrama de dispersión de los datos de densidad de la madera.

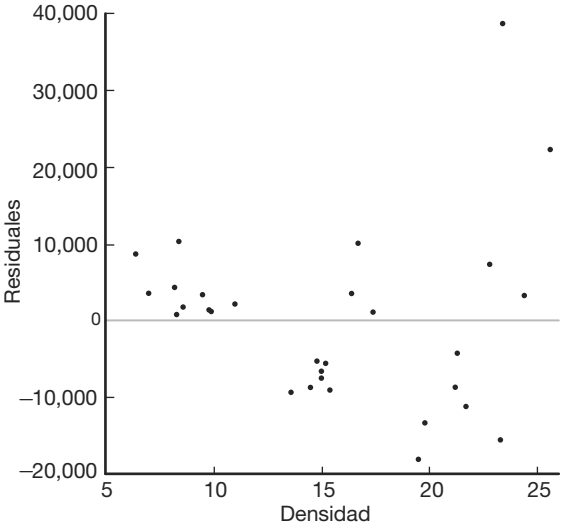


Figura 11.24: Gráfica de los residuales para los datos de densidad de la madera.

y se calcularon los residuales. En la figura 11.24 se presentan los residuales graficados contra las mediciones de la densidad. Difícilmente se trata de un conjunto de residuales ideal o satisfactorio, pues no muestran una distribución aleatoria alrededor del valor de cero. En realidad, los agrupamientos de valores positivos y negativos sugerirían que se debe investigar una tendencia curvilínea en los datos.

Para darnos una idea respecto a la suposición de error normal se dibujó una gráfica de probabilidad normal de los residuales. Es el tipo de gráfica que estudiamos en la sección 8.8, donde el eje horizontal representa la función de distribución normal empírica en una escala que produce una gráfica con línea recta cuando se grafica contra los residuales. En la figura 11.25 se presenta la gráfica de probabilidad normal de los residuales. Esta gráfica no refleja la apariencia de recta que a uno le gustaría ver, lo cual es otro síntoma de una selección errónea, quizá sobresimplificada, de un modelo de regresión.

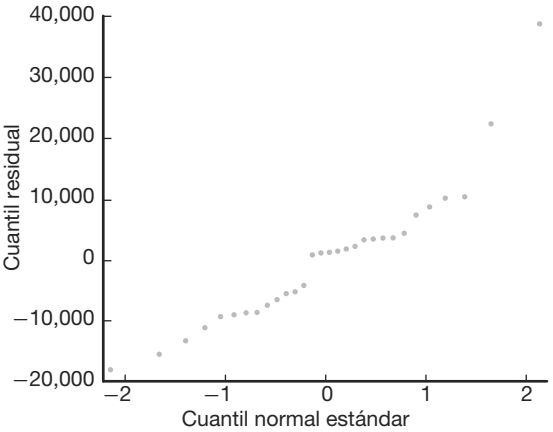


Figura 11.25: Gráfica de probabilidad normal de los residuales para los datos de densidad de la madera.

Los dos tipos de gráficas de residuales y, de hecho, el propio diagrama de dispersión, sugieren que sería adecuado un modelo algo más complicado. Una posibilidad es usar un modelo con transformación de logaritmos naturales. En otras palabras, hay que elegir hacer la regresión de ln y contra x. Esto produce la regresión

$$\widehat{\ln y} = 8\ .257 + 0\ .125x \quad (R^2 = 0.9016).$$

Para darse una idea de si el modelo transformado es más apropiado considere las figuras 11.26 y 11.27, que muestran las gráficas de los residuales de la rigidez [es decir, y_i - antilog $(\widehat{\ln y})$]] en comparación con las de la densidad. La figura 11.26 parece más cercana a un patrón aleatorio alrededor del cero, en tanto que la figura 11.27 con seguridad se acerca más a una línea recta. Esto, además de un valor de R^2 más elevado, sugeriría que el modelo transformado es más apropiado.

11.12 Correlación

Hasta este momento se ha supuesto que la variable regresora independiente x es una variable científica o física en lugar de una variable aleatoria. De hecho, en este contexto es frecuente que x se denomine **variable matemática**, la cual, en el proceso de muestreo, se mide con un error despreciable. En muchas aplicaciones de las técnicas de regresión es más realista suponer que tanto X como Y son variables aleatorias y que las mediciones $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ son observaciones de una población que tiene la función de

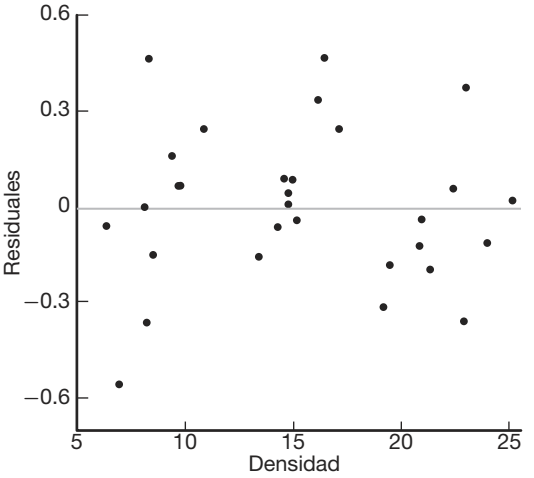


Figura 11.26: Gráfica de residuales donde se utiliza una transformación logarítmica para los datos de densidad de la madera.

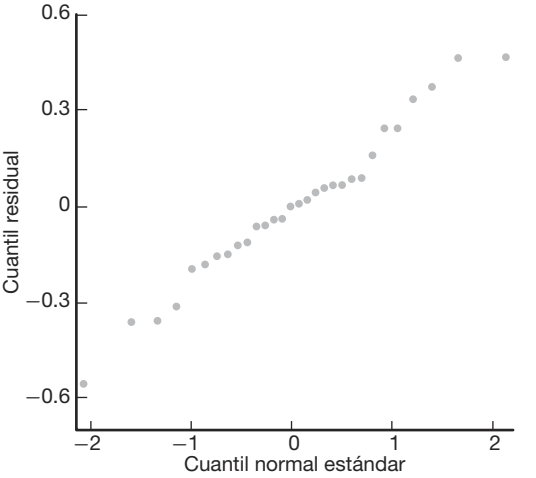


Figura 11.27: Gráfica de probabilidad normal de residuales en la cual se utiliza una transformación logarítmica para los datos de densidad de la madera.

densidad conjunta $f(x, y)$. Debemos considerar el problema de medir la relación entre las dos variables X y Y. Por ejemplo, si X y Y representaran la longitud y la circunferencia de una clase particular de hueso en el cuerpo de un adulto, podríamos realizar un estudio antropológico para determinar si los valores grandes de X se asocian con valores grandes de Y, y viceversa.

Por otro lado, si X representa la antigüedad de un automóvil usado y Y representa su precio de lista al menudeo, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y. El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

En teoría, con frecuencia se supone que la distribución condicional $f(y|x)$ de Y, para valores fijos de X, es normal con media $\mu_{Y|x} = \beta_0 + \beta_1 x$ y varianza $\sigma^2_{Y|x} = \sigma^2$, y que, de igual manera, X se distribuye de forma normal con media μ y varianza σ_x^2. Entonces, la densidad conjunta de X y Y es

$$\begin{aligned} f(x, y) &= n(y|x; \beta_0 + \beta_1 x, \sigma) n(x; \mu_X, \sigma_X) \\ &= \frac{1}{2\pi\sigma_x\sigma} \exp\left\{-\frac{1}{2}\left[\left(\frac{y - \beta_0 - \beta_1 x}{\sigma}\right)^2 + \left(\frac{x - \mu_X}{\sigma_X}\right)^2\right]\right\}, \end{aligned}$$

para $-\infty < [illegible] < \infty$ y $-\infty < y < \infty$.

Escribamos la variable aleatoria Y en la forma

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

donde ahora X es una variable aleatoria independiente del error aleatorio ϵ. Como la media del error aleatorio ϵ es cero, se deduce que

$$\mu_Y = \beta_0 + \beta_1 \mu_X \quad \text{y} \quad \sigma_Y^2 = \sigma^2 + \beta_1^2 \sigma_X^2 .$$

Al sustituir para α y σ^2 en la expresión anterior para $f(x, y)$, se obtiene la **distribución normal bivariada**

$$f(x, y) = \frac{1}{2\pi\sigma_X \sigma_Y \sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$, donde

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \beta_1^2 \frac{\sigma_X^2}{\sigma_Y^2}.$$

La constante ρ (ro) se denomina **coeficiente de correlación de la población** y desempeña un papel importante en muchos problemas de análisis de datos bivariados. Es importante que el lector entienda la interpretación física de este coeficiente de correlación, así como la diferencia entre correlación y regresión. El término *regresión* aún tiene algún significado aquí. De hecho, la línea recta dada por $\mu_{Y|x} = \beta_0 + \beta_1 x$ se sigue llamando recta de regresión, igual que antes, y los estimadores de β_0 y β_1 son idénticos a los que se presentaron en la sección 11.3. El valor de ρ es 0 cuando $\beta_1 = 0$, que resulta cuando en esencia no existe regresión lineal; es decir, cuando la recta de regresión es horizontal y cualquier conocimiento de X es inútil para predecir Y. Como $\sigma_Y^2 \geq \sigma^2$, se debe tener $\rho^2 \leq 1$ y, por lo tanto, $-1 \leq \rho \leq 1$. Los valores de ρ ±1 sólo ocurren cuando $\sigma^2 = 0$, en cuyo caso se tiene una relación lineal perfecta entre las dos variables. Así, un valor de ρ igual a $+1$ implica una relación lineal perfecta con pendiente positiva, en tanto que un valor de ρ igual a -1 resulta de una relación lineal perfecta con pendiente negativa. Entonces, se podría decir que los estimadores muestrales de ρ con magnitud cercana a la unidad implican una buena correlación o **asociación lineal** entre X y Y, mientras que valores cercanos a cero indican poca o ninguna correlación.

Para obtener un estimador muestral de ρ recordemos que en la sección 11.4 aprendimos que la suma de los cuadrados del error es

$$SCE = S_{yy} - b_1 S_{xy}.$$

Al dividir ambos lados de esta ecuación entre S_{yy} y reemplazar S_{xy} con $b_1 S_{xx}$, se obtiene la relación

$$b_1^2 \frac{S_{xx}}{S_{yy}} = 1 - \frac{SCE}{S_{yy}}.$$

El valor de $b_1^2 S_{xx} / S_{yy}$ es igual a cero cuando $b_1 = 0$, lo que ocurrirá cuando los puntos muestrales no tengan relación lineal. Como $S_{yy} \geq SCE$, se concluye que $b_1^2 S_{xx} / S_{xy}$ debe estar entre 0 y 1. En consecuencia, $b_1\sqrt{S_{xx} / S_{yy}}$ debe variar entre -1 y $+1$, y los valores negativos corresponden a rectas con pendientes negativas, mientras que los valores positivos corresponden a rectas con pendientes positivas. Un valor de -1 o $+1$ sucederá cuando $SCE = 0$, pero éste es el caso en el que todos los puntos muestrales caen sobre una línea recta. Por lo tanto, una relación lineal perfecta se da en los datos muestrales cuando $b_1\sqrt{S_{xx} / S_{yy}} = \pm 1$. Es claro que la cantidad $b_1\sqrt{S_{xx} / S_{yy}}$, la cual se designará de aquí en adelante como r, se puede usar como un estimado del coeficiente de correlación ρ de la población. Se acostumbra hacer referencia al estimado r como **coeficiente de correlación producto-momento de Pearson**, o sólo como **coeficiente de correlación muestral**.

Coeficiente de correlación

La medida ρ de la asociación lineal entre dos variables X y Y se estima por medio del **coeficiente de correlación muestral** r, donde

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Hay que tener cuidado en la interpretación de valores de r entre -1 y $+1$. Por ejemplo, valores de r iguales a 0.3 y 0.6 significan sólo que hay dos correlaciones positivas, una un poco más fuerte que la otra. Sería un error concluir que $r = 0.6$ indica una relación lineal dos veces mejor que la del valor $r = 0.3$. Por otro lado, si escribimos

$$r^2 = \frac{S_{xy}^2}{S_{xx}\, S_{yy}} = \frac{SCR}{S_{yy}},$$

entonces r^2, que por lo general se denomina **coeficiente muestral de determinación**, representa la proporción de la variación de S_{yy} explicada por la regresión de Y sobre x, a saber, la SCR. Es decir, r^2 expresa la proporción de la variación total de los valores de la variable Y que son ocasionados o explicados por una relación lineal con los valores de la variable aleatoria X. Así, una correlación de 0.6 significa que 0.36, o 36%, de la variación total de los valores de Y en la muestra se explica mediante la relación lineal con los valores de X.

Ejemplo 11.10: Es importante que los investigadores científicos del área de productos forestales sean capaces de estudiar la correlación entre la anatomía y las propiedades mecánicas de los árboles. Para el estudio *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. Ex Marsh.) and Their Relationships to Mechanical Properties,* realizado por el Departamento de Bosques y Productos Forestales de Virginia Tech, se seleccionaron al azar 29 pinos de Arkansas para investigarlos. En la tabla 11.9 se presentan los datos resultantes sobre la gravedad específica en gramos/cm^3 y el módulo de ruptura en kilopascales (kPa). Calcule e interprete el coeficiente de correlación muestral.

Tabla 11.9: Datos de 29 pinos de Arkansas para el ejemplo 11.10

Gravedad específica, x (g/cm^3)	Módulo de ruptura, y (kPa)	Gravedad específica, x (g/cm^3)	Módulo de ruptura, y (kPa)
0.414	29,186	0.581	85,156
0.383	29,266	0.557	69,571
0.399	26,215	0.550	84,160
0.402	30,162	0.531	73,466
0.442	38,867	0.550	78,610
0.422	37,831	0.556	67,657
0.466	44,576	0.523	74,017
0.500	46,097	0.602	87,291
0.514	59,698	0.569	86,836
0.530	67,705	0.544	82,540
0.569	66,088	0.557	81,699
0.558	78,486	0.530	82,096
0.577	89,869	0.547	75,657
0.572	77,369	0.585	80,490
0.548	67,095		

***Solución*:** A partir de los datos se encuentra que

$$S_{xx} = 0.11273, \quad S_{yy} = 11{,}807{,}324{,}805, \quad S_{xy} = 34{,}422.27572.$$

Por lo tanto,

$$r = \frac{34{,}422.27572}{\sqrt{(0.11273)(11{,}807{,}324{,}805)}} = 0.9435.$$

Un coeficiente de correlación de 0.9435 indica una buena relación lineal entre X y Y. Como $r^2 = 0.8902$, se puede decir que aproximadamente 89% de la variación de los valores de Y es ocasionada por una relación lineal con X.

Una prueba de la hipótesis especial $\rho = 0$ en comparación con una alternativa apropiada es equivalente a probar $\beta_1 = 0$ para el modelo de regresión lineal simple y, por lo tanto, son aplicables los procedimientos de la sección 11.8, donde se usaba la distribución t con $n - 2$ grados de libertad o la distribución F con 1 y $n - 2$ grados de libertad. Sin embargo, si se desea evitar el procedimiento del análisis de varianza y tan sólo calcular el coeficiente de correlación muestral, se podría verificar (véase el ejercicio de repaso 11.66 en la página 438) que el valor t

$$t = \frac{b_1}{s/\sqrt{S_{xx}}}$$

también se puede escribir como

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

que, como antes, es un valor del estadístico T que tiene una distribución t con $n - 2$ grados de libertad.

Ejemplo 11.11: Para los datos del ejemplo 11.10 pruebe la hipótesis de que no existe asociación lineal entre las variables.

***Solución*:** **1.** H_0: $\rho = 0$.
2. H_1: $\rho \neq 0$.
3. $\alpha = 0.05$.
4. Región crítica: $t < -2.052$ o $t > 2.052$.
5. Cálculos: $t = \frac{0.9435\sqrt{27}}{\sqrt{1-0.9435^2}} = 14.79$, $P \approx 0.0001$.
6. Decisión: Rechazar la hipótesis de que no existe asociación lineal.

A partir de la información muestral es fácil efectuar una prueba de la hipótesis más general de que $\rho = \rho_0$ en comparación con una hipótesis alternativa adecuada. Si X y Y siguen una distribución normal bivariada, la cantidad

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

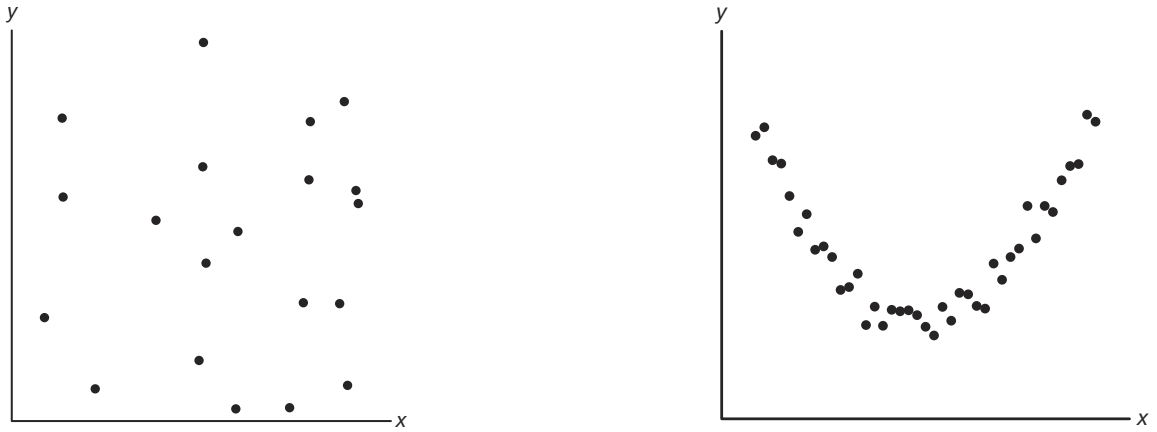
es el valor de una variable aleatoria que sigue aproximadamente la distribución normal con media $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ y varianza $1/(n-3)$. Entonces, el procedimiento de prueba consiste en calcular

$$z = \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$$

y compararlo con los puntos críticos de la distribución normal estándar.

Ejemplo 11.12: Para los datos del ejemplo 11.10 pruebe la hipótesis nula de que $\rho = 0.9$ en comparación con la alternativa de que $\rho > 0.9$. Utilice un nivel de significancia de 0.05.

***Solución*:** **1.** H_0: $\rho = 0.9$.
2. H_1: $\rho > 0.9$.
3. $\alpha = 0.05$.
4. Región crítica: $z > 1.645$.



a) Sin asociación *b*) Relación causal

Figura 11.28: Diagrama de dispersión que muestra correlación de cero.

5. Cálculos:

$$z = \frac{\sqrt{26}}{2} \ln \left[\frac{(1 \text{ } .9435)(0.1)}{(1 - 0.9435)(1.9)} \right] = 1.51, \qquad P = 0.0655.$$

6. Decisión: Existe con certeza alguna evidencia de que el coeficiente de correlación no excede a 0.9. ∎

Debe precisarse que en los estudios de correlación, como en los problemas de regresión lineal, los resultados obtenidos sólo son tan buenos como el modelo que se adopte. En las técnicas de correlación estudiadas aquí se supone que las variables X y Y tienen una densidad normal bivariada, con el valor medio de Y para cada valor de x relacionado en forma lineal con x. Con frecuencia es útil elaborar una gráfica preliminar de los datos experimentales para observar qué tan adecuada es la suposición de linealidad. Un valor del coeficiente de correlación muestral cercano a cero resultará de datos que muestren un efecto estrictamente aleatorio, como los de la figura 11.28*a*, lo que implica que hay poca o ninguna relación causal. Es importante recordar que el coeficiente de correlación entre dos variables es una medida de su relación lineal, y que un valor de $r = 0$ implica *falta de linealidad y no falta de asociación*. Por lo tanto, si existiera una relación cuadrática fuerte entre X y Y, como la que se observa en la figura 11.28*b*, aún se podría obtener una correlación de cero que indicaría una relación no lineal.

Ejercicios

11.43 Calcule e interprete el coeficiente de correlación para las siguientes calificaciones de 6 estudiantes seleccionados al azar:

Calificación en matemáticas	70	92	80	74	65	83
Calificación en inglés	74	84	63	87	78	90

11.44 Remítase al ejercicio 11.1 de la página 398 y suponga que x y y son variables aleatorias con una distribución normal bivariada:

a) Calcule r.

b) Pruebe la hipótesis de que $\rho = 0$ en comparación con la alternativa de que $\rho \neq 0$ a un nivel de significancia de 0.05.

11.45 Remítase al ejercicio 11.13 de la página 400, suponga una distribución normal bivariada para x y y.

a) Calcule r.

b) Pruebe la hipótesis nula de que $\rho = -0.5$, en comparación con la alternativa de que $\rho < -0.5$, a un nivel de significancia de 0.025.

c) Determine el porcentaje de la variación en la cantidad de partículas eliminadas que se debe a cambios en la cantidad de lluvia diaria.

11.46 En el ejercicio 11.43 pruebe la hipótesis de que $\rho = 0$ en comparación con la alternativa de que $\rho \neq 0$. Utilice un nivel de significancia de 0.05.

11.47 Los datos siguientes se obtuvieron en un estudio de la relación entre el peso y el tamaño del pecho de niños al momento de nacer.

Peso (kg)	Tamaño del pecho (cm)
2.75	29.5
2.15	26.3
4.41	32.2
5.52	36.5
3.21	27.2
4.32	27.7
2.31	28.3
4.30	30.3
3.71	28.7

a) Calcule r.

b) Pruebe la hipótesis nula de que $\rho = 0$ en comparación con la alternativa de que $\rho > 0$ a un nivel de significancia de 0.01.

c) ¿Qué porcentaje de la variación del tamaño del pecho de los niños es explicado por la diferencia de peso?

Ejercicios de repaso

11.48 Remítase al ejercicio 11.8 de la página 399 y construya

a) un intervalo de confianza de 95% para la calificación promedio en el curso de los estudiantes que obtuvieron 35 puntos en el examen de colocación;

b) un intervalo de predicción de 95% para la calificación del curso de un estudiante que obtuvo 35 puntos en el examen de colocación.

11.49 El Centro de Consulta Estadística de Virginia Tech analizó datos sobre las marmotas normales para el Departamento de Veterinaria. Las variables de interés fueron el peso corporal en gramos y el peso del corazón en gramos. Se deseaba desarrollar una ecuación de regresión lineal con el fin de determinar si había una relación lineal significativa entre el peso del corazón y el peso total del cuerpo.

Peso corporal (gramos)	Peso del corazón (gramos)
4050	11.2
2465	12.4
3120	10.5
5700	13.2
2595	9.8
3640	11.0
2050	10.8
4235	10.4
2935	12.2
4975	11.2
3690	10.8
2800	14.2
2775	12.2
2170	10.0
2370	12.3
2055	12.5
2025	11.8
2645	16.0
2675	13.8

Utilice el peso del corazón como la variable independiente, el peso del cuerpo como la dependiente y haga un ajuste de regresión lineal simple con los siguientes datos. Además, pruebe la hipótesis de que H_0: $\beta_1 = 0$ en comparación con H_1: $\beta_1 \neq 0$. Saque conclusiones.

11.50 A continuación se presentan las cantidades de sólidos eliminados de cierto material cuando se expone a periodos de secado de diferentes duraciones.

x (horas)	y (gramos)	
4.4	13.1	14.2
4.5	9.0	11.5
4.8	10.4	11.5
5.5	13.8	14.8
5.7	12.7	15.1
5.9	9.9	12.7
6.3	13.8	16.5
6.9	16.4	15.7
7.5	17.6	16.9
7.8	18.3	17.2

a) Estime la recta de regresión lineal.

b) Pruebe si es adecuado el modelo lineal a un nivel de significancia de 0.05.

11.51 Remítase al ejercicio 11.9 de la página 399 y construya

a) un intervalo de confianza de 95% para las ventas semanales promedio cuando se gastan \$45 en publicidad.

b) un intervalo de predicción de 95% para las ventas semanales cuando se gastan \$45 en publicidad.

11.52 Se diseñó un experimento para el Departamento de Ingeniería de Materiales de Virginia Tech con el fin de estudiar las propiedades de deterioro del nitrógeno con base en las mediciones de la presión de hidrógeno

electrolítico. Se utilizó una solución al 0.1 *N* NaOH y el material era cierto tipo de acero inoxidable. La densidad de corriente de carga catódica fue controlada y variada en cuatro niveles. Se observó la presión de hidrógeno efectiva como la respuesta. A continuación se presentan los datos.

Ensayo	Densidad de corriente de carga, x (mA/cm^2)	Presión de hidrógeno efectiva, y (atm)
1	0.5	86.1
2	0.5	92.1
3	0.5	64.7
4	0.5	74.7
5	1.5	223.6
6	1.5	202.1
7	1.5	132.9
8	2.5	413.5
9	2.5	231.5
10	2.5	466.7
11	2.5	365.3
12	3.5	493.7
13	3.5	382.3
14	3.5	447.2
15	3.5	563.8

a) Efectúe un análisis de regresión lineal simple de y con x.

b) Calcule la suma de cuadrados del error puro y haga una prueba para la falta de ajuste.

c) ¿La información del inciso *b* indica la necesidad de un modelo en x más allá de una regresión de primer orden? Explique su respuesta.

11.53 Los datos siguientes representan la calificación en química de una muestra aleatoria de 12 estudiantes de nuevo ingreso a cierta universidad, así como sus calificaciones en una prueba de inteligencia aplicada mientras estudiaban el último año de preparatoria.

Estudiante	Calificación en la prueba, x	Calificación en química, y
1	65	85
2	50	74
3	55	76
4	65	90
5	55	85
6	70	87
7	65	94
8	70	98
9	55	81
10	70	91
11	50	76
12	55	74

a) Calcule e interprete el coeficiente de correlación de la muestra.

b) Establezca las suposiciones necesarias acerca de las variables aleatorias.

c) Pruebe la hipótesis de que $\rho = 0.5$ en comparación con la alternativa de que $\rho > 0.5$. Use un valor P para las conclusiones.

11.54 La sección de negocios del *Washington Times* de marzo de 1997 listaba 21 diferentes computadoras e impresoras usadas, así como sus precios de lista. También se listaba la oferta promedio. En la figura 11.29 de la página 439 se presenta una parte de los resultados impresos por computadora del análisis de regresión usando el programa *SAS*.

a) Explique la diferencia entre el intervalo de confianza sobre la media y el intervalo de predicción.

b) Explique por qué los errores estándar de la predicción varían de una observación a otra.

c) ¿Cuál observación tiene el menor error estándar de la predicción? Explique su respuesta.

11.55 Considere los datos de los vehículos de *Consumer Reports* que se incluyen en la figura 11.30 de la página 440. El peso se indica en toneladas, el rendimiento en millas por galón y también se incluye el cociente de manejo. Se ajustó un modelo de regresión que relaciona el peso x con el rendimiento y. En la figura 11.30 de la página 440 se observa una salida parcial del *SAS* con los resultados de dicho análisis de regresión, y en la figura 11.31 de la página 441 se incluye una gráfica de los residuales y el peso de cada vehículo.

a) A partir del análisis y la gráfica de los residuales, ¿se podría concluir que cabría la posibilidad de encontrar un modelo mejorado si se usara una transformación? Explique su respuesta.

b) Ajuste el modelo reemplazando el peso con el logaritmo del peso. Comente los resultados.

c) Ajuste un modelo reemplazando mpg con los galones por cada 100 millas recorridas, como se reporta con frecuencia el rendimiento del combustible en otros países. ¿Cuál de los tres modelos es preferible? Explique su respuesta.

11.56 A continuación se presentan las observaciones registradas del producto de una reacción química tomadas a temperaturas diferentes:

x (°C)	y (%)	x (°C)	y (%)
150	75.4	150	77.7
150	81.2	200	84.4
200	85.5	200	85.7
250	89.0	250	89.4
250	90.5	300	94.8
300	96.7	300	95.3

a) Grafique los datos.

b) ¿La gráfica indica que la relación es lineal?

c) Haga un análisis de regresión lineal simple y pruebe la falta de ajuste.

d) Saque conclusiones con base en el resultado del inciso *c*.

11.57 La prueba de acondicionamiento físico es un aspecto importante del entrenamiento atlético. Una medida común para determinar la aptitud cardiovascular es el volumen máximo de oxígeno que se inhala al realizar un ejercicio extenuante. Se realizó un estudio con 24 hombres de mediana edad para analizar cómo el tiempo que les tomaba correr una distancia de dos millas influía en el oxígeno que consumían, el cual se midió con métodos estándar de laboratorio mientras los sujetos se ejercitaban en una banda sin fin. El trabajo fue publicado en el artículo "Maximal Oxygen Intake Prediction in Young and Middle Aged Males", *Journal of Sports Medicine* **9**, 1969, 17-22. A continuación se presentan los datos.

Sujeto	*y*, Volumen máximo de O_2	*x*, Tiempo en segundos
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	46.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.67	700
23	60.62	748
24	56.73	775

a) Estime los parámetros en un modelo de regresión lineal simple.

b) ¿El tiempo que toma correr dos millas influye de forma significativa en la cantidad máxima de oxígeno consumido? Utilice H_0: $\beta_0 = 0$ en comparación con H_1: $\beta_1 \neq 0$.

c) Grafique los residuales en una gráfica en comparación con x y haga comentarios sobre qué tan apropiado es el modelo lineal simple.

11.58 Suponga que cierto científico postula el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, \ldots, n,$$

y β_0 es un **valor conocido** no necesariamente igual a cero.

a) ¿Cuál es el estimador apropiado de mínimos cuadrados de β_1? Justifique su respuesta.

b) ¿Cuál es la varianza del estimador de la pendiente?

11.59 Para el modelo de regresión lineal simple demuestre que $E(s^2) = \sigma^2$.

11.60 Suponga que las ϵ_i son independientes y que se distribuyen normalmente con medias de cero y varianza común σ^2, y demuestre que B_0, el estimador de mínimos cuadrados de β_0 en $\mu_{Y|x} = \beta_0 + \beta_1 x$, se distribuye de manera normal con media β_0 y varianza

$$\sigma^2_{B_0} = \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sigma^2.$$

11.61 Para un modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{i} = 1, 2, \ldots, n,$$

donde las ϵ_i son independientes y se distribuyen normalmente con medias de cero y varianzas iguales σ^2, demuestre que

$$B_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

y tienen covarianza de cero.

11.62 Demuestre, en el caso de un ajuste de mínimos cuadrados al modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

que $\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$.

11.63 Considere la situación del ejercicio de repaso 11.62 pero suponga que $n = 2$, es decir, que sólo disponemos de dos puntos de datos. Argumente que la recta de regresión de mínimos cuadrados tendrá como resultado $(y_1 - \hat{y}_1) = (y_2 - \hat{y}_2) = 0$. También demuestre que para este caso $R^2 = 1.0$.

11.64 En el ejercicio de repaso 11.62 se pidió al estudiante que demostrara que $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ para un modelo de regresión lineal simple estándar. ¿Se cumple también para un modelo con intersección en el origen? Demuestre su respuesta, ya sea afirmativa o negativa.

11.65 Suponga que un experimentador plantea un modelo como

$$Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

cuando en realidad una variable adicional, digamos x_2, también contribuye linealmente a la respuesta. Entonces, el verdadero modelo es dado por

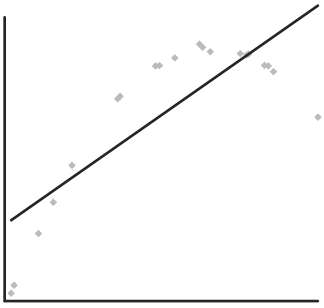
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Calcule el valor esperado del estimador

$$B_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)Y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}.$$

11.66 Demuestre los pasos necesarios para convertir la ecuación $r = \frac{b_1}{s/\sqrt{S_{xx}}}$ a la forma equivalente $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

11.67 Considere el siguiente grupo ficticio de datos, donde la línea que los atraviesa representa la recta de regresión lineal simple ajustada. Grafique los residuales.



11.68 **Proyecto**: Este proyecto se puede realizar en grupos o de manera individual. Cada grupo o persona debe encontrar un grupo de datos, preferiblemente de su campo de estudios, aunque también pueden ser de otro campo. Los datos se deben ajustar al esquema de regresión, con una variable de regresión x y una variable de respuesta y. Determine con cuidado cuál variable es x y cuál es y. Tal vez necesite consultar una revista científica de su campo si no cuenta con otros datos experimentales.

a) Grafique y contra x. Comente sobre la relación que se observa en la gráfica.

b) Diseñe un modelo de regresión adecuado a partir de los datos. Utilice una regresión lineal simple o ajuste un modelo polinomial a los datos. Comente acerca de medidas de calidad.

c) Grafique los residuales como se indica en el texto. Verifique posibles violaciones de los supuestos. Muestre de forma gráfica una representación de los intervalos de confianza de una respuesta media graficada en comparación con x. Haga comentarios al respecto.

R-Square	Coeff Var	Root MSE	Price Mean
0.967472	7.923338	70.83841	894.0476

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	59.93749137	38.34195754	1.56	0.1345
Buyer	1.04731316	0.04405635	23.77	<.0001

product	Buyer	Price	Value	Predict	Std Err Predict	Lower 95% Mean	Upper 95% Mean	Lower 95% Predict	Upper 95% Predict
IBM PS/1 486/66 420MB	325	375	400.31	25.8906		346.12	454.50	242.46	558.17
IBM ThinkPad 500	450	625	531.23	21.7232		485.76	576.70	376.15	686.31
IBM Think-Dad 755CX	1700	1850	1840.37	42.7041		1750.99	1929.75	1667.25	2013.49
AST Pentium 90 540MB	800	875	897.79	15.4590		865.43	930.14	746.03	1049.54
Dell Pentium 75 1GB	650	700	740.69	16.7503		705.63	775.75	588.34	893.05
Gateway 486/75 320MB	700	750	793.06	16.0314		759.50	826.61	641.04	945.07
Clone 586/133 1GB	500	600	583.59	20.2363		541.24	625.95	429.40	737.79
Compaq Contura 4/25 120MB	450	600	531.23	21.7232		485.76	576.70	376.15	686.31
Compaq Deskpro P90 1.2GB	800	850	897.79	15.4590		865.43	930.14	746.03	1049.54
Micron P75 810MB	800	675	897.79	15.4590		865.43	930.14	746.03	1049.54
Micron P100 1.2GB	900	975	1002.52	16.1176		968.78	1036.25	850.46	1154.58
Mac Quadra 840AV 500MB	450	575	531.23	21.7232		485.76	576.70	376.15	686.31
Mac Performer 6116 700MB	700	775	793.06	16.0314		759.50	826.61	641.04	945.07
PowerBook 540c 320MB	1400	1500	1526.18	30.7579		1461.80	1590.55	1364.54	1687.82
PowerBook 5300 500MB	1350	1575	1473.81	28.8747		1413.37	1534.25	1313.70	1633.92
Power Mac 7500/100 1GB	1150	1325	1264.35	21.9454		1218.42	1310.28	1109.13	1419.57
NEC Versa 486 340MB	800	900	897.79	15.4590		865.43	930.14	746.03	1049.54
Toshiba 1960CS 320MB	700	825	793.06	16.0314		759.50	826.61	641.04	945.07
Toshiba 4800VCT 500MB	1000	1150	1107.25	17.8715		1069.85	1144.66	954.34	1260.16
HP Laserjet III	350	475	426.50	25.0157		374.14	478.86	269.26	583.74
Apple LaserWriter Pro 63	750	800	845.42	15.5930		812.79	878.06	693.61	997.24

Figura 11.29: Salida por computadora de los resultados del *SAS* que presenta el análisis parcial de datos del ejercicio de repaso 11.54.

Obs	Model	WT	MPG	DR_RATIO
1	Buick Estate Wagon	4.360	16.9	2.73
2	Ford Country Squire Wagon	4.054	15.5	2.26
3	Chevy Ma libu Wagon	3.605	19.2	2.56
4	Chrysler LeBaron Wagon	3.940	18.5	2.45
5	Chevette	2.155	30.0	3.70
6	Toyota Corona	2.560	27.5	3.05
7	Datsun 510	2.300	27.2	3.54
8	Dodge Omni	2.230	30.9	3.37
9	Audi 5000	2.830	20.3	3.90
10	Volvo 240 CL	3.140	17.0	3.50
11	Saab 99 GLE	2.795	21.6	3.77
12	Peugeot 694 SL	3.410	16.2	3.58
13	Buick Century Special	3.380	20.6	2.73
14	Mercury Zephyr	3.070	20.8	3.08
15	Dodge Aspen	3.620	18.6	2.71
16	AMC Concord D/L	3.410	18.1	2.73
17	Chevy Caprice Classic	3.840	17.0	2.41
18	Ford LTP	3.725	17.6	2.26
19	Mercury Grand Marquis	3.955	16.5	2.26
20	Dodge St Regis	3.830	18.2	2.45
21	Ford Mustang 4	2.585	26.5	3.08
22	Ford Mustang Ghia	2.910	21.9	3.08
23	Macda GLC	1.975	34.1	3.73
24	Dodge Colt	1.915	35.1	2.97
25	AMC Spirit	2.670	27.4	3.08
26	VW Scirocco	1.990	31.5	3.78
27	Honda Accord LX	2.135	29.5	3.05
28	Buick Skylark	2.570	28.4	2.53
29	Chevy Citation	2.595	28.8	2.69
30	Olds Omega	2.700	26.8	2.84
31	Pontiac Phoenix	2.556	33.5	2.69
32	Plymouth Horizon	2.200	34.2	3.37
33	Datsun 210	2.020	31.8	3.70
34	Fiat Strada	2.130	37.3	3.10
35	VW Dasher	2.190	30.5	3.70
36	Datsun 810	2.815	22.0	3.70
37	BMW 320i	2.600	21.5	3.64
38	VW Rabbit	1.925	31.9	3.78

R-Square	Coeff Var	Root MSE	MPG Mean
0.817244	11.46010	2.837580	24.76053

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	48.67928080	1.94053995	25.09	<.0001
WT	-8.36243141	0.65908398	-12.69	<.0001

Figura 11.30: Salida de computadora de los resultados del *SAS* que muestra el análisis parcial de los datos del ejercicio de repaso 11.55.

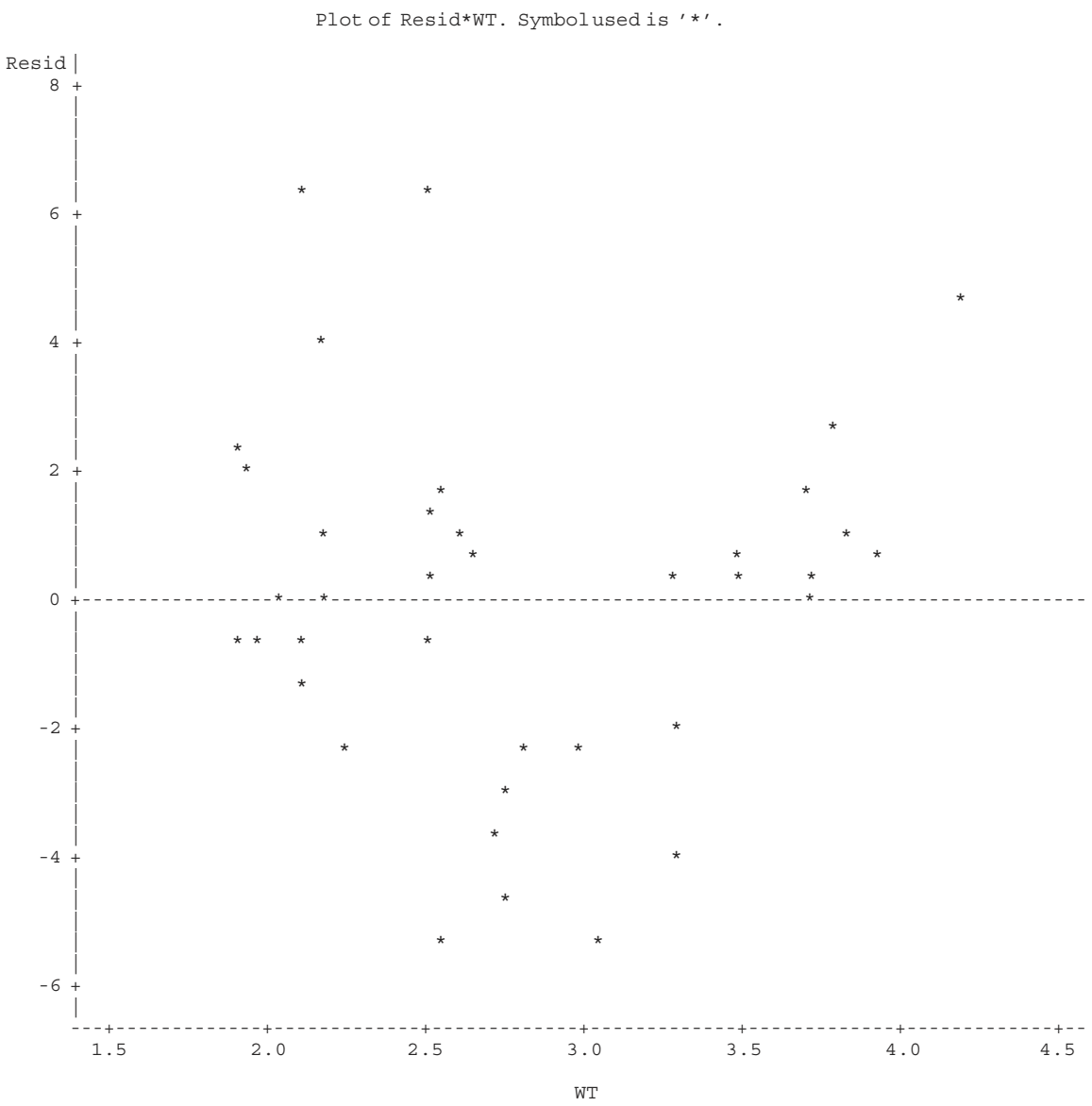


Figura 11.31: Salida de computadora de los resultados del *SAS* que muestra la gráfica de residuales del ejercicio de repaso 11.55.

11.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Cada vez que se considere utilizar la regresión lineal simple no sólo es recomendable elaborar una gráfica de los datos, sino esencial. Siempre es edificante elaborar una gráfica de los residuales ordinarios y otra de la probabilidad normal de los mismos. Además, en el capítulo 12 se presentará e ilustrará un tipo adicional de residual en forma estandarizada. Todas esas gráficas están diseñadas para detectar la transgresión de las suposiciones.

El uso de los estadísticos t para las pruebas sobre los coeficientes de regresión es razonablemente robusto para la suposición de normalidad. La suposición de varianza homogénea es crucial y las gráficas de los residuales están diseñadas para detectar una violación.

El material de este capítulo se utiliza ampliamente en los capítulos 12 a 15. Toda la información acerca del método de los mínimos cuadrados para la elaboración de modelos de regresión se utilizará en el capítulo 12. La diferencia es que en ese capítulo se abordan las condiciones científicas en las que hay más de una sola variable x, es decir, más de una variable de regresión. Sin embargo, también utilizaremos el material de este capítulo en el que se exponen los diagnósticos de regresión, los tipos de gráficas residuales, las medidas de la calidad del modelo, etcétera. El estudiante notará que en el capítulo 12 habrá más complicaciones, lo cual se debe a que los problemas de los modelos de regresión múltiple suelen incluir el fundamento de las cuestiones respecto a cómo las diversas variables de regresión entran en el modelo, e incluso el tema de cuáles variables deben permanecer en el modelo. De hecho, el capítulo 15 incluye el uso constante de los modelos de regresión, pero en el resumen al final del capítulo 12 presentaremos una vista preliminar de la conexión.

Capítulo 12

Regresión lineal múltiple y ciertos modelos de regresión no lineal

12.1 Introducción

En la mayoría de los problemas de investigación en los que se aplica el análisis de regresión se necesita más de una variable independiente para el modelo de regresión. La complejidad de la mayoría de mecanismos científicos es tal que, con el fin de predecir una respuesta importante, se requiere un **modelo de regresión múltiple**. Cuando un modelo es lineal en los coeficientes se denomina **modelo de regresión lineal múltiple**. Para el caso de k variables independientes, el modelo que da x_1, x_2,..., x_k, la media de $Y|x_1$, x_2,..., x_k es el modelo de regresión lineal múltiple

$$\mu_{Y|x_1, x_2, ..., x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

y la respuesta estimada se obtiene a partir de la ecuación de regresión muestral

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

donde cada coeficiente de regresión β_i se estima por medio de b_i, a partir de los datos muestrales, usando el método de los mínimos cuadrados. Como ocurre en el caso de una sola variable independiente, a menudo el modelo de regresión lineal múltiple es una representación adecuada de una estructura más complicada dentro de ciertos rangos de las variables independientes.

También se pueden aplicar técnicas similares de mínimos cuadrados para estimar los coeficientes cuando el modelo lineal incluye, por ejemplo, potencias y productos de las variables independientes. Un ejemplo de esto se presentaría cuando $k = 1$, en cuyo caso el experimentador podría pensar que las medias $\mu_{Y|x}$ no caen sobre una línea recta, sino que se describen de manera más adecuada mediante el **modelo de regresión polinomial**

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

y la respuesta estimada se obtiene de la ecuación de regresión polinomial

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

En ocasiones se genera confusión al decir que un modelo polinomial es un modelo lineal. Sin embargo, los estadísticos normalmente se refieren a un modelo lineal como aquel en el que los parámetros ocurren en forma lineal, independientemente de cómo las variables independientes entran en el modelo. Un ejemplo de modelo no lineal es la **relación exponencial**

$$\mu_{Y|x} = \alpha\beta^x,$$

que se estima mediante la ecuación de regresión

$$\hat{y} = ab^x.$$

En ciencias e ingeniería hay muchos fenómenos cuya naturaleza no es inherentemente lineal y, cuando se conoce su verdadera estructura, no hay duda de que habría que intentar ajustar el modelo real. Existe mucha literatura acerca de la estimación de modelos no lineales por medio de mínimos cuadrados. Los modelos no lineales que se analizan en este capítulo se relacionan con condiciones no ideales, en las cuales el analista está seguro de que la respuesta y, por lo tanto, el error de respuesta del modelo no se distribuyen normalmente sino que, más bien, tienen una distribución binomial o de Poisson. Estas situaciones ocurren a menudo en la práctica.

El estudiante que busque profundizar en la explicación de la regresión no lineal debe consultar la obra de Myers *Classical and Modern Regression with Applications* (1990; véase la bibliografía).

12.2 Estimación de los coeficientes

En esta sección se calculan los estimadores de mínimos cuadrados de los parámetros β_0, β_1,..., β_k mediante el ajuste del modelo de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

a los puntos de los datos

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i); \quad i = 1, 2, \ldots, n \text{ y } n > k\},$$

donde y_i es la respuesta observada a los valores x_{1i}, x_{2i},..., x_{ki} de las k variables independientes x_1, x_2,..., x_k. Se supone que cada observación (x_{1i}, x_{2i},..., x_{ki}, y_i) satisface la siguiente ecuación:

Modelo de regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

o bien,

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

donde ϵ_i y e_i son el error aleatorio y el residual, respectivamente, asociados con la respuesta y_i y con el valor ajustado $\hat{y}_i$.

Como en el caso de la regresión lineal simple, se supone que los ϵ_i son independientes y están distribuidos en forma idéntica con media cero y varianza común σ^2.

Si usamos el concepto de mínimos cuadrados para obtener los estimados b_0, b_1,..., b_k, minimizamos la expresión

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

Si, a su vez, diferenciamos la *SCE* respecto a b_0, b_1,..., b_k e igualamos el resultado a cero, generamos el conjunto de $k + 1$ **ecuaciones normales para la regresión lineal múltiple**.

Ecuaciones normales de estimación para la regresión lineal múltiple

$$
\begin{aligned}
nb_0 + b_1\sum_{i=1}^{n} x_{1i} \quad &+ b_2\sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k\sum_{i=1}^{n} x_{ki} \quad = \sum_{i=1}^{n} y_i \\
b_0\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 \quad &+ b_2\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
&\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \qquad\qquad \vdots \\
b_0\sum_{i=1}^{n} x_{ki} + b_1\sum_{i=1}^{n} x_{ki}x_{1i} &+ b_2\sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{ki}^2 \quad = \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}
$$

Estas ecuaciones se pueden resolver para b_0, b_1, b_2,..., b_k utilizando cualquier método apropiado que permita resolver sistemas de ecuaciones lineales. Casi todos los programas estadísticos de cómputo se pueden utilizar para obtener soluciones numéricas de las ecuaciones anteriores.

Ejemplo 12.1: Se sometió a prueba un grupo de camiones ligeros con motores que utilizan diesel como combustible para saber si la humedad, la temperatura del aire y la presión barométrica influyen en la cantidad de óxido nitroso que emiten (en ppm). Las emisiones se midieron en distintos momentos y en diversas condiciones experimentales. Los datos se presentan en la tabla 12.1. El modelo es

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

o, en forma equivalente,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

Ajuste este modelo de regresión lineal múltiple a los datos con los que cuenta y luego estime la cantidad de óxido nitroso que emiten los camiones en las siguientes condiciones: 50% de humedad, temperatura de 76°F y una presión barométrica de 29.30.

Tabla 12.1: Datos para el ejemplo 12.1

Óxido nitroso, y	Humedad, x_1	Temp., x_2	Presión, x_3	Óxido nitroso, y	Humedad, x_1	Temp., x_2	Presión, x_3
0.90	72.4	76.3	29.18	1.07	23.2	76.8	29.38
0.91	41.6	70.3	29.35	0.94	47.4	86.6	29.35
0.96	34.3	77.1	29.24	1.10	31.5	76.9	29.63
0.89	35.1	68.0	29.27	1.10	10.6	86.3	29.56
1.00	10.7	79.0	29.78	1.10	11.2	86.0	29.48
1.10	12.9	67.4	29.39	0.91	73.3	76.3	29.40
1.15	8.3	66.8	29.69	0.87	75.4	77.9	29.28
1.03	20.1	76.9	29.48	0.78	96.6	78.7	29.29
0.77	72.2	77.7	29.09	0.82	107.4	86.8	29.03
1.07	24.0	67.7	29.60	0.95	54.9	70.9	29.37

Fuente: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions", EPA-600/2-77-116. U. S. Environmental Protection Agency.

***Solución*:** La solución del conjunto de las ecuaciones de estimación produce los estimadores únicos

$$b_0 = -3.507778,\ b_1 = -0.002625,\ b_2 = 0.000799,\ b_3 = 0.154155.$$

Por lo tanto, la ecuación de regresión es

$$\hat{y} = -3.507778 - 0.002625x_1 + 0\ .000799x_2 + 0.154155x_3.$$

Para 50% de humedad, una temperatura de 76°F y una presión barométrica de 29.30, la cantidad estimada de óxido nitroso emitido es

$$\begin{aligned}\hat{y} =& -3.507778 - 0.002625(50.0) + 0\ .000799(76.0) + 0\ .1541553(29.30)\\ =& \ 0.9384 \text{ ppm.}\end{aligned}$$

Regresión polinomial

Ahora suponga que se desea ajustar la ecuación polinomial

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

a los n pares de observaciones $\{(x_i, y_i);\ i = 1, 2,..., n\}$. Cada observación, y_i, satisface la ecuación

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

o bien,

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

donde r es el grado del polinomio y ϵ_i y e_i son, de nuevo, el error aleatorio y el residual asociados con la respuesta y_i y con el valor ajustado $\hat{y}_i$, respectivamente. Aquí el número de pares, n, debe ser al menos $r + 1$, que es el número de parámetros por estimar.

Observe que el modelo polinomial se puede considerar un caso especial del modelo de regresión lineal múltiple más general, donde establecemos $x_1 = x$, $x_2 = x^2$,..., $x_r = x^r$. Las ecuaciones normales adoptan la misma forma que las que aparecen en la página 445. Luego se resuelven para b_0, b_1, b_2,..., b_r.

Ejemplo 12.2: Dados los datos

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

ajuste una curva de regresión de la forma $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$, luego, estime $\mu_{Y|2}$.

***Solución*:** A partir de los datos se encuentra que

$$\begin{aligned}10b_0 + \quad 45b_1 + \quad 285\, b_2 &= 63.7,\\ 45b_0 + \quad 285b_1 + \quad 2025\, b_2 &= 307.3,\\ 285b_0 + 2025\, b_1 + 15{,}333b_2 &= 2153.3.\end{aligned}$$

Al resolver las ecuaciones normales se obtiene

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

Por lo tanto,

$$\hat{y} = 8.698 - 2.341x + 0.288x^2.$$

Cuando $x = 2$ el estimado de $\mu_{Y|2}$ es

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

Ejemplo 12.3: Los datos de la tabla 12.2 representan el porcentaje de impurezas que resultaron de diversas temperaturas y del tiempo de esterilización durante una reacción asociada con la fabricación de cierta bebida. Estime los coeficientes de regresión en el modelo polinomial

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

para $i = 1, 2,...,18$.

Tabla 12.2: Datos para el ejemplo 12.3

Tiempo de esterilización, x_2 (min)	**Temperatura, x_1 (°C)**		
	75	**100**	**125**
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

***Solución*:** Si usamos las ecuaciones normales, obtenemos

$$b_0 = 56.4411, \quad b_1 = -0.36190, \quad b_2 = -2.75299,$$
$$b_{11} = 0.00081, \quad b_{22} = 0.08173, \quad b_{12} = 0.00314,$$

y nuestra ecuación de regresión estimada es

$$\hat{y} = 56.4411 - 0.36190x_1 - 2.75299x_2 + 0.00081x_1^2 + 0.08173x_2^2 + 0.00314x_1x_2.$$

Muchos de los principios y procedimientos asociados con la estimación de funciones de regresión polinomiales caen en la categoría de **metodología de respuesta superficial**, que es un conjunto de técnicas que los científicos e ingenieros de muchos campos han utilizado con bastante éxito. Las x_i^2 se denominan **términos cuadráticos puros** y las $x_i x_j$ $(i \neq j)$ se conocen como **términos de interacción**. Dichas técnicas a menudo se aplican a problemas tales como seleccionar un diseño experimental adecuado, en particular en casos en los que un número muy grande de variables entra en el modelo; y elegir condiciones óptimas de operación para $x_1, x_2,..., x_k$. Para profundizar en este tema se recomienda al lector consultar la obra de Myers, Montgomery y Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (2009; véase la bibliografía).

12.3 Modelo de regresión lineal en el que se utilizan matrices

Al ajustar un modelo de regresión lineal múltiple, en particular cuando contiene más de dos variables, tener conocimientos sobre la teoría de matrices facilita considerablemente el manejo de las matemáticas. Suponga que el experimentador tiene k variables

independientes $x_1, x_2, ..., x_k$ y n observaciones $y_1, y_2, ..., y_n$, cada una de las cuales se puede expresar con la ecuación

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

Este modelo representa en esencia a n ecuaciones que describen cómo se generan los valores de la respuesta durante el proceso científico. Si usamos la notación de matrices, podemos escribir la ecuación siguiente

Modelo lineal general

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon},$$

donde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Después, el método de mínimos cuadrados para la estimación de β, que se estudió en la sección 12.2, implica calcular **b**, para lo cual

$$SCE = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

se minimiza. Este proceso de minimización implica resolver para **b** en la ecuación

$$\frac{\partial}{\partial \mathbf{b}}(SCE) = \mathbf{0}.$$

Aquí no presentaremos los detalles respecto a cómo se resuelven las ecuaciones anteriores. El resultado se reduce a la solución de **b** en

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

Observe la naturaleza de la matriz **X**. Además del elemento inicial, el i-ésimo renglón representa los valores de x que dan lugar a la respuesta y_i. Si escribimos

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

y

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

nos permite escribir las ecuaciones normales en la forma de matriz

$$\mathbf{Ab} = \mathbf{g}.$$

Si la matriz $\mathbf{A}$ es no singular, la solución para los coeficientes de regresión se escribe como

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

De esta manera, obtenemos la ecuación de predicción o regresión resolviendo un conjunto de $k + 1$ ecuaciones con un número igual de incógnitas. Esto implica el invertir la matriz $\mathbf{X}'\mathbf{X}$ de orden $k + 1$ por $k + 1$. En la mayoría de libros que tratan sobre determinantes y matrices elementales se explican las técnicas para invertir matrices. Por supuesto, existen muchos paquetes de cómputo veloces para resolver problemas de regresión múltiple, los cuales no sólo proporcionan estimados de los coeficientes de regresión, sino que también ofrecen otra clase de información relevante para hacer inferencias acerca de la ecuación de regresión.

Ejemplo 12.4: Se midió el porcentaje de supervivencia de los espermatozoides de cierto tipo de semen animal, después de almacenarlo con distintas combinaciones de concentraciones de tres materiales que se emplean para incrementar la supervivencia. En la tabla 12.3 se presentan los datos. Obtenga el modelo de regresión lineal múltiple para los datos.

Tabla 12.3: Datos para el ejemplo 12.4

y (% de supervivencia)	x_1 (peso %)	x_2 (peso %)	x_3 (peso %)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

***Solución*:** Las ecuaciones de estimación por mínimos cuadrados, $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$, son

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}.$$

A partir de una salida de computadora se obtienen los elementos de la matriz inversa

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix},$$

y, luego, utilizando la relación $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, se obtienen los siguientes coeficientes de regresión estimados

$$b_0 = 39.1574,\ b_1 = 1.0161,\ b_2 = -1.8616,\ b_3 = -0.3433.$$

Entonces, la ecuación de regresión estimada es

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3.$$

Ejercicios

12.1 Se llevó a cabo un conjunto de ensayos experimentales con un horno para determinar una forma de predecir el tiempo de cocción, y, a diferentes niveles de ancho del horno, x_1, y a diferentes temperaturas, x_2. Se registraron los siguientes datos:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

12.2 En *Applied Spectroscopy* se estudiaron las propiedades de reflectancia infrarroja de un líquido viscoso que se utiliza como lubricante en la industria electrónica. El experimento que se diseñó consistió en medir el efecto de frecuencia de banda, x_1, y el espesor de película, x_2, sobre la densidad óptica, y, usando un espectrómetro infrarrojo Perkin-Elmer Modelo 621. (*Fuente:* Pacansky, J., England, C. D. y Wattman, R., 1986).

y	x_1	x_2
0.231	740	1.10
0.107	740	0.62
0.053	740	0.31
0.129	805	1.10
0.069	805	0.62
0.030	805	0.31
1.005	980	1.10
0.559	980	0.62
0.321	980	0.31
2.948	1235	1.10
1.633	1235	0.62
0.934	1235	0.31

Estime la ecuación de regresión lineal múltiple

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

12.3 En el ejercicio de repaso 11.53 de la página 437 suponga que también se proporciona el número de periodos de clase perdidos por los 12 estudiantes que toman el curso de química. A continuación se presentan los datos completos.

Estudiante	Calificación en química, y	Calificación en el examen, x_1	Clases perdidas, x_2
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

a) Ajuste una ecuación de regresión lineal múltiple de la forma $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2$.

b) Estime la calificación de química para un estudiante que en la prueba de inteligencia obtuvo 60 de calificación y perdió 4 clases.

12.4 Se realizó un experimento para determinar si era posible predecir el peso de un animal después de un periodo determinado con base en su peso inicial y la cantidad de alimento que consumía. Se registraron los siguientes datos, en kilogramos:

Peso final, y	Peso inicial, x_1	Peso del alimento, x_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

a) Ajuste una ecuación de regresión múltiple de la forma

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

b) Prediga cuánto pesará un animal que comienza pesando 35 kilogramos después de consumir 250 kilogramos de alimento.

12.5 Se cree que la energía eléctrica que una planta química consume cada mes se relaciona con la temperatura ambiental promedio, x_1, el número de días del mes, x_2, la pureza promedio del producto, x_3, y las toneladas fabricadas del producto, x_4. Se dispone de datos históricos del año anterior, los cuales se presentan en la siguiente tabla.

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

a) Ajuste un modelo de regresión lineal múltiple usando el conjunto de datos anterior.

b) Prediga el consumo de energía para un mes en que $x_1 = 75^\circ$F, $x_2 = 24$ días, $x_3 = 90\%$ y $x_4 = 98$ toneladas.

12.6 Se realizó un experimento sobre un modelo nuevo de una marca de automóvil específica para determinar la distancia de frenado a distintas velocidades. Se registraron los siguientes datos.

Velocidad, v (km/h)	35	50	65	80	95	110
Distancia de frenado, d (m)	16	26	41	62	88	119

a) Ajuste una curva de regresión múltiple de la forma $\mu_{D|v} = \beta_0 + \beta_1 v + \beta_2 v^2$.

b) Estime la distancia de frenado cuando el automóvil viaja a 70 kilómetros por hora.

12.7 Se realizó un experimento con el fin de determinar si el flujo sanguíneo cerebral de los seres humanos se podía predecir a partir de la tensión arterial del oxígeno (milímetros de mercurio). En el estudio participaron 15 pacientes y se reunieron los siguientes datos:

Flujo sanguíneo, y	**Tensión arterial del oxígeno, x**
84.33	603.40
87.80	582.50
82.20	556.20
78.21	594.60
78.44	558.90
80.01	575.20
83.53	580.10
79.46	451.20
75.22	404.00
76.58	484.00
77.90	452.40
78.80	448.40
80.67	334.80
86.60	320.30
78.20	350.30

Estime la ecuación de regresión cuadrática

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

12.8 El siguiente es un conjunto de datos experimentales codificados acerca de la resistencia a la compresión de una aleación específica para distintos valores de la concentración de cierto aditivo:

Concentración, x	**Resistencia a la compresión, y**		
10.0	25.2	27.3	28.7
15.0	29.8	31.1	27.8
20.0	31.2	32.6	29.7
25.0	31.7	30.1	32.3
30.0	29.4	30.8	32.8

a) Estime la ecuación de regresión cuadrática $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$.

b) Pruebe la falta de ajuste del modelo.

12.9 *a*) Ajuste una ecuación de regresión múltiple de la forma $\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x^2$ para los datos del ejemplo 11.8 de la página 420.

b) Estime el producto de la reacción química para una temperatura de 225°C.

12.10 Para los datos siguientes

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

a) Ajuste el modelo cúbico $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

b) Prediga el valor de Y cuando $x = 2$.

12.11 Se realizó un experimento para estudiar el tamaño de los calamares consumidos por tiburones y atunes. Las variables regresoras son características de la boca del calamar. Los datos del estudio son los siguientes:

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.01	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49
1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

En el estudio las variables regresoras y la respuesta considerada son

x_1 = longitud del rostral, en pulgadas,
x_2 = longitud de la aleta, en pulgadas,
x_3 = longitud del rostral a la cola, en pulgadas,
x_4 = longitud de la cola a la aleta, en pulgadas,
x_5 = ancho, en pulgadas,
y = peso, en libras.

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1, x_2, x_3, x_4, x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

12.12 Los siguientes datos reflejan información obtenida en 17 hospitales de la marina estadounidense ubicados en diversos sitios del mundo. Los regresores son variables de la carga de trabajo, es decir, conceptos que dan como resultado la necesidad de personal en un hospital. A continuación se presenta una descripción breve de las variables:

y = horas de trabajo mensuales,
x_1 = carga diaria promedio de pacientes,
x_2 = exposiciones de rayos X mensuales,
x_3 = días-cama ocupados por mes,
x_4 = población elegible en el área/1000,
x_5 = duración promedio de la estancia de un paciente, en días.

Sitio	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1003.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11,520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20,106	3655.08	180.5	6.15	3503.93
11	96.00	13,313	2912.00	60.9	5.88	3571.59
12	131.42	10,771	3921.00	103.7	4.88	3741.40
13	127.21	15,543	3865.67	126.8	5.50	4026.52
14	252.90	36,194	7684.10	157.7	7.00	10,343.81
15	409.20	34,703	12,446.33	169.4	10.75	11,732.17
16	463.70	39,204	14,098.40	331.4	7.05	15,414.94
17	510.22	86,533	15,524.00	371.6	6.35	18,854.45

El objetivo es generar una ecuación empírica para estimar (o predecir) las necesidades de personal en los hospitales de la marina. Calcule la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1, x_2, x_3, x_4, x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

12.13 Se llevó a cabo un estudio sobre un tipo de conexión para conocer la relación entre la cantidad de desgaste, y, para x_1 = viscosidad del aceite, y x_2 = carga. Se obtuvieron los datos siguientes. (Tomado de *Response Surface Methodology*, Myers, Montgomery y Anderson-Cook, 2009).

y	x_1	x_2	y	x_1	x_2
193	1.6	851	230	15.5	816
172	22.0	1058	91	43.0	1201
113	33.0	1357	125	40.0	1115

a) Estime los parámetros desconocidos de la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

b) Prediga el desgaste cuando la viscosidad del aceite sea de 20 y la carga sea de 1200.

12.14 Once estudiantes normalistas participaron en un programa de evaluación diseñado para medir la eficacia de los maestros y determinar cuáles factores son importantes. La medición de la respuesta consistió en una evaluación cuantitativa del maestro. Las variables regresoras fueron las calificaciones de cuatro pruebas estandarizadas aplicadas a cada maestro. Los datos son los siguientes:

y	x_1	x_2	x_3	x_4
410	69	125	59.00	55.66
569	57	131	31.75	63.97
425	77	141	80.50	45.32
344	81	122	75.00	46.67
324	0	141	49.00	41.21
505	53	152	49.35	43.83
235	77	141	60.75	41.61
501	76	132	41.25	64.57
400	65	157	50.75	42.41
584	97	166	32.25	57.95
434	76	141	54.50	57.90

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1, x_2, x_3, x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

12.15 Con el fin de determinar la relación entre la calificación de su desempeño laboral (y) y las calificaciones en cuatro exámenes, el departamento de personal de cierta empresa industrial realizó un estudio en el que participaron 12 sujetos. Los datos son los siguientes:

y	x_1	x_2	x_3	x_4
11.2	56.5	71.0	38.5	43.0
14.5	59.5	72.5	38.2	44.8
17.2	69.2	76.0	42.5	49.0
17.8	74.5	79.5	43.4	56.3
19.3	81.2	84.0	47.5	60.2
24.5	88.0	86.2	47.4	62.0
21.2	78.2	80.5	44.5	58.1
16.9	69.0	72.0	41.8	48.1
14.8	58.1	68.0	42.1	46.0
20.0	80.5	85.0	48.1	60.3
13.2	58.3	71.0	37.5	47.1
22.5	84.0	87.2	51.0	65.2

Estime los coeficientes de regresión del modelo

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4.$$

12.16 Un ingeniero de una empresa de semiconductores desea modelar la relación entre la ganancia o hFE de un dispositivo (y) y tres parámetros: RS del emisor (x_1), RS de la base (x_2) y RS del emisor a la base (x_3). A continuación se muestran los datos:

x_1, RS del emisor	x_2, RS de la base	x_3, E-B-RS	y, hFE
14.62	226.0	7.000	128.40
15.63	220.0	3.375	52.62
14.62	217.4	6.375	113.90
15.00	220.0	6.000	98.01
14.50	226.5	7.625	139.90
15.25	224.1	6.000	102.60
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.60
15.50	217.6	5.000	82.68
15.13	228.5	6.625	112.60
15.50	230.2	5.750	97.52
16.12	226.5	3.750	59.06
15.13	226.6	6.125	111.80
15.63	225.6	5.375	89.09
15.38	234.0	8.875	171.90
15.50	230.0	4.000	66.80
14.25	224.3	8.000	157.10
14.50	240.5	10.870	208.40
14.62	223.7	7.375	133.40

(Datos de Myers, Montgomery y Anderson-Cook, 2009).

a) Ajuste una regresión lineal múltiple para los datos.
b) Prediga hFE cuando $x_1 = 14$, $x_2 = 220$ y $x_3 = 5$.

12.4 Propiedades de los estimadores de mínimos cuadrados

Las medias y varianzas de los estimadores b_0, b_1,..., b_k se obtienen con facilidad si se hacen ciertas suposiciones sobre los errores aleatorios ϵ_1, ϵ_2,..., ϵ_k, que son idénticas a las que se hacen en el caso de la regresión lineal simple. Si suponemos que dichos errores son independientes, cada uno con media igual a cero y varianza σ^2, entonces podemos demostrar que b_0, b_1,..., b_k son, respectivamente, estimadores no sesgados de los coeficientes de regresión β_0, β_1,..., β_k. Además, las varianzas de las b se obtienen por medio de los elementos del inverso de la matriz **A**. Observe que los elementos fuera de la diagonal de $\mathbf{A} = \mathbf{X}'\mathbf{X}$ representan sumas de productos de los elementos en las columnas de **X**; mientras que los elementos en la diagonal de **A** son las sumas de los cuadrados de los elementos en las columnas de **X**. La matriz inversa, $\mathbf{A^{-1}}$, aparte del multiplicador σ^2, representa la **matriz de varianza-covarianza** de los coeficientes de regresión estimados. Es decir, los elementos de la matriz $\mathbf{A^{-1}\sigma^2}$ muestran las varianzas de b_0, b_1,..., b_k en la diagonal principal y las covarianzas fuera de la diagonal. Por ejemplo, en un problema de regresión lineal múltiple con $k = 2$ se podría escribir

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

con los elementos debajo de la diagonal principal determinados por la simetría de la matriz. Entonces, se escribe

$$\sigma_{b_i}^2 = c_{ii}\,\sigma^2, \qquad i = 0, 1, 2,$$
$$\sigma_{b_i b_j} = \text{Cov}(b_i, b_j) = c_{ij}\,\sigma^2, \quad i \neq j.$$

Desde luego, los estimados de las varianzas y, por lo tanto, sus errores estándar, se obtienen reemplazando σ^2 con el estimado apropiado, el cual se obtuvo a partir de los datos experimentales. Un estimado no sesgado de σ^2 de nuevo se define en términos de

la suma de cuadrados del error, que se calcula utilizando la fórmula establecida en el teorema 12.1. En el teorema las suposiciones se basan en los ϵ_i descritos con anterioridad.

Teorema 12.1: Para la ecuación de regresión lineal

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon},$$

un estimador insesgado de σ^2 es dado por el error o media cuadrática residual

$$s^2 = \frac{SCE}{n-k-1}, \quad \text{donde} \quad SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Podemos ver que, para el caso de la regresión lineal simple, el teorema 12.1 representa una generalización del teorema 11.1. La prueba se deja como ejercicio para el lector. Al igual que en el caso de la regresión lineal más simple, el estimado de s^2 es una medida de la variación de los errores de la predicción o residuales. En las secciones 12.10 y 12.11 se presentan otras inferencias importantes relacionadas con la ecuación ajustada de regresión, con base en los valores de los residuales individuales $e_i = y_i - \hat{y}_i$, $i = 1, 2, ..., n$.

La suma de cuadrados del error y de la regresión adoptan la misma forma y desempeñan el mismo papel que en el caso de la regresión lineal simple. De hecho, la identidad de la suma de cuadrados

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

se sigue cumpliendo, y se conserva la notación anterior, que es,

$$STCC = SCR + SCE,$$

con

$$STCC = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{suma de cuadrados total}$$

y

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{suma de cuadrados de regresión}$$

Hay k grados de libertad asociados con la *SCR*, y, como siempre, la *STCC* tiene $n - 1$ grados de libertad. Por lo tanto, después de restar, la *SCE* tiene $n - k - 1$ grados de libertad. Así, nuestro estimado de σ^2 de nuevo es dado por la suma de cuadrados del error dividida entre sus grados de libertad. Las tres sumas de cuadrados aparecen en la salida de resultados de la mayoría de los programas de cómputo de regresión múltiple. Observe que la condición $n > k$ en la sección 12.2 garantiza que los grados de libertad de la *SCE* no sean negativos.

Análisis de varianza en la regresión múltiple

La partición de la suma total de cuadrados en sus componentes, la suma de cuadrados de regresión y del error desempeña un papel importante. Puede efectuarse un **análisis de varianza** que arroje luz sobre la calidad de la ecuación de regresión. Una hipótesis que sirve para determinar si el modelo explica una cantidad significativa de variación, es la siguiente:

$$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0\,.$$

El análisis de varianza implica una prueba F, mediante una tabla, como la siguiente:

Fuente	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	SCR	k	$CMR = \frac{SCR}{k}$	$f = \frac{CMR}{CME}$
Error	SCE	$n-(k+1)$	$CME = \frac{SCE}{n-(k+1)}$	
Total	$STCC$	$n-1$		

Se trata de una **prueba de cola superior**. El rechazo de H_0 significa que la **ecuación de regresión difiere de una constante**. Es decir, al menos una variable regresora es importante. En las secciones que siguen se estudia más el uso del análisis de varianza.

Otra utilidad del cuadrado medio del error (o cuadrado medio residual) estriba en su uso para la prueba de hipótesis y la estimación de intervalos de confianza que se estudian en la sección 12.5. Además, el cuadrado medio del error desempeña un papel importante en las situaciones en las que el científico busca el mejor modelo entre un conjunto de ellos que están en competencia. Muchos criterios de construcción de modelos incluyen el estadístico s^2. En la sección 12.11 se presentan criterios para comparar modelos en competencia.

12.5 Inferencias en la regresión lineal múltiple

El conocimiento de la distribución de los estimadores del coeficiente individual facilita al experimentador construir intervalos de confianza para los coeficientes y hacer pruebas de hipótesis acerca de ellos. Recuerde que en la sección 12.4 estudiamos que b_j ($j = 0$, 1, 2,..., k) se distribuyen de forma normal con media β_j y varianza $c_{jj}\sigma^2$. De esta manera, se puede utilizar el estadístico

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$$

con $n - k - 1$ grados de libertad para probar hipótesis y construir intervalos de confianza sobre β_j. Por ejemplo, si queremos probar

$$H_0\colon\ \beta_j = \beta_{j0},$$
$$H_1\colon\ \beta_j \neq \beta_{j0},$$

se calcula el estadístico t anterior y no se rechaza H_0 si $-t_{\alpha/2} < t < t_{\alpha/2}$, donde $t_{\alpha/2}$ tiene $n - k - 1$ grados de libertad.

Ejemplo 12.5: Para el modelo del ejemplo 12.4 pruebe la hipótesis de que $\beta_2 = -2.5$ en comparación con la alternativa de que $\beta_2 > -2.5$ a un nivel de significancia de 0.05.

***Solución*:**

$$H_0: \ \beta_2 = -2.5,$$
$$H_1: \ \beta_2 > -2.5.$$

Cálculos:

$$t = \frac{b_2 - \beta_{20}}{s\sqrt{c_{22}}} = \frac{-1.8616 + 2.5}{2.073\sqrt{0.0166}} = 2.390,$$
$$P = P(T > 2.390) = 0.04.$$

Decisión: Rechazar H_0 y concluir que $\beta_2 > -2.5$.

Pruebas *t* individuales para la selección de variables

La prueba t que se utiliza con más frecuencia en la regresión múltiple es aquella que prueba la importancia de los coeficientes individuales, es decir, H_0: $\beta_j = 0$ en comparación con la hipótesis alternativa H_1: $\beta_j \neq 0$. Con frecuencia estas pruebas contribuyen a lo que se denomina **selección de variables**, con la cual el analista intenta llegar al modelo más útil, es decir, a la elección de cuál regresor utilizar. Aquí debemos destacar que, si se encuentra que un coeficiente es insignificante, es decir, si **no se rechaza** la hipótesis H_0: $\beta_j = 0$, la conclusión que se obtiene es que la **variable** es insignificante (explica una cantidad insignificante de la variación de y) **en la presencia de los demás regresores del modelo**. Más adelante se profundizará en este punto.

Inferencias sobre la respuesta media y la predicción

Una de las inferencias más útiles que se pueden hacer con respecto a la calidad de la respuesta predicha y_0, correspondiente a los valores x_{10}, x_{20},..., x_{k0}, es el intervalo de confianza sobre la respuesta media $\mu_{y|\, x_{10}, x_{20},...,x_{k0}}$. Estamos interesados en construir un intervalo de confianza sobre la respuesta media para el conjunto de condiciones determinadas por

$$\mathbf{x}_0' = [1, x_{10}, x_{20}, \ldots, x_{k0}].$$

Se aumentan en 1 las condiciones sobre las x para facilitar la notación de matrices. La normalidad en los ϵ_i producen normalidad en los b_j, y la media y la varianza siguen siendo las mismas, como se indica en la sección 12.4. Así es la covarianza entre b_i y b_j para $i \neq j$. De esta manera,

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_{j0}$$

también se distribuye normalmente y es, de hecho, un estimador no sesgado para la **respuesta media** sobre la que se intenta ligar un intervalo de confianza. La varianza de $\hat{y}_0$, escrita con notación de matriz simplemente como función de σ^2, $(\mathbf{X}'\mathbf{X})^{-1}$, y el vector de condiciones, $\mathbf{x}_0'$ es

$$\sigma^2_{\hat{y}_0} = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0.$$

Si esta expresión se extendiera para un caso dado, por ejemplo $k = 2$, ya vimos que explica de manera apropiada la varianza de b_j y la covarianza de b_i y b_j, para $i \neq j$. Después de sustituir σ^2 con s^2, según se plantea en el teorema 12.1, el intervalo de confianza del $100(1 - \alpha)\%$ se puede construir sobre $\mu_{Y|x} = x_{10},x_{20},...,x_{k0}$ a partir del estadístico

$$T = \frac{\hat{y}_0 - \mu_{Y|x_{10},\ x_{20},...,\ x_{k0}}}{s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

que tiene una distribución t con $n - k - 1$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_{10}, x_{20}, ..., x_{k0}}$

Un intervalo de confianza de $100(1 - \alpha)\%$ para la **respuesta media** $\mu_{Y|x_{10}, x_{20}, ..., x_{k0}}$ es

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_{Y|x_{10}, x_{20}, ..., x_{k0}} < \hat{y}_0 + t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - k - 1$ grados de libertad.

Es frecuente que a la cantidad $s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ se le denomine **error estándar de la predicción** y aparece en la salida de resultados de muchos paquetes de cómputo para regresión.

Ejemplo 12.6: Con los datos del ejemplo 12.4 construya un intervalo de confianza de 95% para la respuesta media, cuando $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$.

***Solución*:** De la ecuación de regresión del ejemplo 12.4, el porcentaje estimado de supervivencia cuando $x_1 = 3\%$, $x_2 = 8\%$, y $x_3 = 9\%$, es:

$$\hat{y} = 39.1574 + (1.0161)(3) - (1.8616)(8) - (0.3433)(9) = 24.2232.$$

Y luego se determina que

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = [1, 3, 8, 9]\begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 8 \\ 9 \end{bmatrix}$$

$$= 0.1267.$$

Si utilizamos el cuadrado medio del error, $s^2 = 4.298$ o $s = 2.073$, y la tabla A.4, observamos que $t_{0.025} = 2.262$ para 9 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para el porcentaje medio de supervivencia para $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$ es dado por

$$24.2232 - (2.262)(2.073)\sqrt{0.1267} < \mu_{Y|3,8,9} < 24.2232 + (2.262)(2.073)\sqrt{0.1267},$$

o simplemente $22.5541 < \mu_{Y|3,8,9} < 25.8923$. ■

Como ocurre en el caso de la regresión lineal simple, necesitamos distinguir con claridad entre el intervalo de confianza sobre la respuesta media y el intervalo de predicción sobre una *respuesta observada*. Esta última proporciona un límite dentro del cual podemos decir que, con un grado preseleccionado de certidumbre, caerá una respuesta nueva observada.

Nuevamente se establece un intervalo de predicción para una sola respuesta predicha y_0 al considerar la diferencia $\hat{y}_0 - y_0$. Se puede demostrar que la distribución del muestreo es normal con media

$$\mu_{\hat{y}_0 - y_0} = 0$$

y varianza

$$\sigma^2_{\hat{y}_0 - y_0} = \sigma^2[1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0].$$

Por consiguiente, se puede construir un intervalo de predicción del $100(1 - \alpha)\%$ para un solo valor de predicción y_0 a partir del estadístico

$$T = \frac{\hat{y}_0 - y_0}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

el cual tiene una distribución t con $n - k - 1$ grados de libertad.

Intervalo de predicción para y_0

Un intervalo de predicción del $100(1 - \alpha)\%$ para una **sola respuesta** y_0 es dado por

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < y_0 < \hat{y}_0 + t_{\alpha/2}s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - k - 1$ grados de libertad.

Ejemplo 12.7: Con los datos del ejemplo 12.4 construya un intervalo de predicción de 95% para una respuesta individual del porcentaje de supervivencia, cuando $x_1 = 3\%$, $x_2 = 8\%$, y $x_3 = 9\%$.

***Solución*:** Si nos remitimos a los resultados del ejemplo 12.6, encontramos que el intervalo de predicción de 95% para la respuesta y_0, cuando $x_1 = 3\%$, $x_2 = 8\%$, y $x_3 = 9\%$, es

$$24.2232 - (2.262)(2.073)\sqrt{1.1267} < y_0 < 24.2232 + (2.262)(2.073)\sqrt{1.1267},$$

que se reduce a $19.2459 < y_0 < 29.2005$. Observe que, como se esperaba, el intervalo de predicción es considerablemente más ancho que el intervalo de confianza para el porcentaje medio de supervivencia del ejemplo 12.6. ■

Salida de resultados comentado para los datos del ejemplo 12.4

La figura 12.1 muestra una salida de resultados por computadora con comentarios para el ajuste de regresión lineal múltiple de los datos del ejemplo 12.4. Se empleó el paquete *SAS*.

Observe los estimados de los parámetros del modelo, los errores estándar y los estadísticos t que aparecen en el listado. Los errores estándar se calcularon a partir de las raíces cuadradas de los elementos de la diagonal $(\mathbf{X}'\mathbf{X})^{-1}s^2$. En dicha ilustración la variable x_3 es insignificante en presencia de x_1 y x_2 con base en la prueba t y el valor P correspondiente de 0.5916. Los términos CLM y CLI son intervalos de confianza sobre la respuesta media y los límites de predicción sobre una observación individual, respectivamente. La prueba f en el análisis de varianza indica que se explica una cantidad significativa de variabilidad. Como ejemplo de las interpretaciones de CLM y CLI, considere la observación 10. Con una observación de 25.2000 y un valor predicho de 26.0676 tenemos 95% de confianza en que la respuesta media está entre 24.5024 y 27.6329, y en que una observación nueva caerá entre 21.1238 y 31.0114 con una probabilidad de 0.95. El valor R^2 de 0.9117 implica que el modelo explica el 91.17% de la variabilidad de la respuesta. En la sección 12.6 se analiza más a fondo R^2.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	399.45437	133.15146	30.98	<.0001
Error	9	38.67640	4.29738		
Corrected Total	12	438.13077			

Root MSE	2.07301	R-Square	0.9117
Dependent Mean	29.03846	Adj R-Sq	0.8823
Coeff Var	7.13885		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39.15735	5.88706	6.65	<.0001
x1	1	1.01610	0.19090	5.32	0.0005
x2	1	-1.86165	0.26733	-6.96	<.0001
x3	1	-0.34326	0.61705	-0.56	0.5916

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	25.5000	27.3514	1.4152	24.1500	30.5528	21.6734	33.0294	-1.8514
2	31.2000	32.2623	0.7846	30.4875	34.0371	27.2482	37.2764	-1.0623
3	25.9000	27.3495	1.3588	24.2757	30.4234	21.7425	32.9566	-1.4495
4	38.4000	38.3096	1.2818	35.4099	41.2093	32.7960	43.8232	0.0904
5	18.4000	15.5447	1.5789	11.9730	19.1165	9.6499	21.4395	2.8553
6	26.7000	26.1081	1.0358	23.7649	28.4512	20.8658	31.3503	0.5919
7	26.4000	28.2532	0.8094	26.4222	30.0841	23.2189	33.2874	-1.8532
8	25.9000	26.2219	0.9732	24.0204	28.4233	21.0414	31.4023	-0.3219
9	32.0000	32.0882	0.7828	30.3175	33.8589	27.0755	37.1008	-0.0882
10	25.2000	26.0676	0.6919	24.5024	27.6329	21.1238	31.0114	-0.8676
11	39.7000	37.2524	1.3070	34.2957	40.2090	31.7086	42.7961	2.4476
12	35.7000	32.4879	1.4648	29.1743	35.8015	26.7459	38.2300	3.2121
13	26.5000	28.2032	0.9841	25.9771	30.4294	23.0122	33.3943	-1.7032

Figura 12.1: Salida de resultados del *SAS* para los datos del ejemplo 12.4.

Más sobre el análisis de varianza en la regresión múltiple (opcional)

En la sección 12.4 se estudió brevemente la partición de la suma total de cuadrados $\sum_{i=1}^{n}(y_i-\bar{y})^2$ en sus dos componentes, el modelo de regresión y la suma de cuadrados del error (que se ilustran en la figura 12.1). El análisis de varianza conduce a la prueba de

$$H_0:\ \beta_1=\beta_2=\beta_3=\cdots=\beta_k=0.$$

El rechazo de la hipótesis nula implica una interpretación importante para el científico o el ingeniero. (A quienes les interese profundizar en el tema del uso de matrices les será útil estudiar el desarrollo de estas sumas de cuadrados que se usan en el ANOVA).

En primer lugar, de la sección 12.3 recuerde que **b**, el vector de los estimadores de mínimos cuadrados, es dado por

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Una partición de la **suma de cuadrados no corregida**,

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2$$

en dos componentes es dada por

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{b}'\mathbf{X}'\mathbf{y} + (\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}) \\ &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + [\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}].\end{aligned}$$

El segundo término (entre corchetes) en el lado derecho es tan sólo la suma de cuadrados del error $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. El lector debería observar que una expresión alternativa para la suma de cuadrados del error es

$$SCE = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

El término $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ se denomina **suma de cuadrados de la regresión**. Sin embargo, no se trata de la expresión $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ que se usó para probar la "importancia" de los términos $b_1, b_2, ..., b_k$, sino más bien de

$$\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \hat{y}_i^2,$$

que es la suma de cuadrados de la regresión no corregida para la media. Como tal, sólo se podría usar para probar si la ecuación de regresión difiere significativamente de cero, es decir,

$$H_0: \ \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

En general, esto no es tan importante como probar

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

dado que esto plantea que la respuesta media es una constante, no necesariamente cero.

Grados de libertad

Así, la partición de las sumas de cuadrados y los grados de libertad se reduce a

Fuente	Suma de cuadrados	gl
Regresión	$\sum_{i=1}^{n} \hat{y}_i^2 = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$k + 1$
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$	$n - (k + 1)$
Total	$\sum_{i=1}^{n} y_i^2 = \mathbf{y}'\mathbf{y}$	n

Hipótesis de interés

Desde luego, la hipótesis de interés para un ANOVA debe eliminar el papel de la intersección según se describió anteriormente. En términos estrictos, si H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, entonces la recta de regresión estimada es simplemente $\hat{y}_i = \bar{y}$. Como resultado, en realidad se busca evidencia de que la ecuación de regresión "varíe a partir de una constante". Así, la suma de cuadrados total y la suma de regresión deben corregirse para la media. Como resultado, tenemos

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

En notación de matrices esto es simplemente

$$\mathbf{y}'[\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y} = \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y} + \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

En esta expresión el **1** sólo es un vector de n unos. Como resultado, simplemente restamos

$$\mathbf{y}'\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

de $\mathbf{y}'\mathbf{y}$ y de $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, es decir, corrigiendo la suma de cuadrados total y la de regresión para la media.

Por último, la partición apropiada de las sumas de cuadrados con grados de libertad es como sigue:

Fuente	Suma de cuadrados	gl
Regresión	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}]\mathbf{y}$	k
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$	$n - (k + 1)$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$	$n - 1$

Ésta es la tabla ANOVA que aparece en la salida de resultados por computadora de la figura 12.1. Es frecuente denominar a la expresión $\mathbf{y}'[\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$ como la **suma de cuadrados de la regresión asociada con la media**, y se le asigna 1 grado de libertad.

Ejercicios

12.17 Para los datos del ejercicio 12.2 de la página 450, estime σ^2.

12.18 Para los datos del ejercicio 12.1 de la página 450, estime σ^2.

12.19 Para los datos del ejercicio 12.5 de la página 450, estime σ^2.

12.20 Obtenga estimados de las varianzas y la covarianza de los estimadores b_1 y b_2, del ejercicio 12.2 de la página 450.

12.21 Remítase al ejercicio 12.5 de la página 450 y obtenga estimados de

a) $\sigma^2_{b_2}$;
b) $\text{Cov}(b_1, b_4)$.

12.22 Para el modelo del ejercicio 12.7 de la página 451, a un nivel de significancia de 0.05 pruebe la hipótesis de que $\beta_2 = 0$, en comparación con la hipótesis alternativa de que $\beta_2 \neq 0$.

12.23 Para el modelo del ejercicio 12.2 de la página 450 a un nivel de significancia de 0.05, pruebe la hipótesis de que $\beta_1 = 0$, en comparación con la hipótesis alternativa de que $\beta_1 \neq 0$.

12.24 Para el modelo del ejercicio 12.1 de la página 450 pruebe la hipótesis de que $\beta_1 = 2$, en comparación con la hipótesis alternativa de que $\beta_1 \neq 2$. Utilice un valor *P* en sus conclusiones.

12.25 Utilice los datos del ejercicio 12.2 de la página 450 y el estimado de σ^2 del ejercicio 12.17 para calcular intervalos de confianza de 95% para la respuesta predicha y la respuesta media cuando $x_1 = 900$ y $x_2 = 1.00$.

12.26 Para el ejercicio 12.8 de la página 451 construya un intervalo de confianza de 90% para la resistencia media a la compresión cuando la concentración es $x = 19.5$ y se utiliza un modelo cuadrático.

12.27 Utilice los datos del ejercicio 12.5 de la página 450 y el estimado de σ^2 del ejercicio 12.19 para calcular intervalos de confianza de 95% para la respuesta predicha y la respuesta media cuando $x_1 = 75$, $x_2 = 24$, $x_3 = 90$ y $x_4 = 98$.

12.28 Considere los siguientes datos del ejercicio 12.13 de la página 452.

y (desgaste)	x_1 (viscosidad del aceite)	x_2 (carga)
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

a) Estime σ^2 usando regresión múltiple de *y* sobre x_1 y x_2.

b) Calcule valores predichos, un intervalo de confianza de 95% para el desgaste promedio y un intervalo de predicción de 95% para el desgaste observado si $x_1 = 20$ y $x_2 = 1000$.

12.29 Con los datos del ejercicio 12.28, y a un nivel de 0.05, pruebe:

a) H_0: $\beta_1 = 0$ en comparación con H_1: $\beta_1 = 0$;

b) H_0: $\beta_2 = 0$ en comparación con H_1: $\beta_2 = 0$.

c) ¿Existe alguna razón para creer que habría que cambiar el modelo del ejercicio 12.28? Explique su respuesta.

12.30 Utilice los datos del ejercicio 12.16 de la página 453.

a) Estime σ^2 usando la regresión múltiple de *y* sobre x_1, x_2 y x_3;

b) Calcule un intervalo de predicción de 95% para la ganancia observada con los tres regresores en $x_1 = 15.0$, $x_2 = 220.0$ y $x_3 = 6.0$.

12.6 Selección de un modelo ajustado mediante la prueba de hipótesis

En muchas situaciones de regresión los coeficientes individuales revisten importancia para el experimentador. Por ejemplo, en una aplicación de economía, $\beta_1, \beta_2, \ldots$ podrían tener un significado en particular, por lo que el economista tendría un interés especial en los intervalos de confianza y en las pruebas de hipótesis sobre dichos parámetros. Sin embargo, considere una situación de química industrial en la que el modelo propuesto supone que el producto de la reacción depende linealmente de la temperatura y concentración de la reacción de cierto catalizador. Es probable que se sepa que éste no es el verdadero modelo, sino una aproximación adecuada; de manera que el interés no estribaría en los parámetros individuales, sino en la capacidad de la función en su conjunto para predecir la respuesta verdadera en el rango de las variables consideradas. Por lo tanto, en esta situación, se pondría más énfasis en $\sigma^2_{\hat{Y}}$, los intervalos de confianza de la respuesta media, y así sucesivamente, y disminuiría el interés en las inferencias sobre los parámetros individuales.

El experimentador que utiliza análisis de regresión también está interesado en eliminar variables cuando la situación impone que, además de llegar a una ecuación de pronóstico funcional, debe encontrar la "mejor regresión" que implique sólo variables que sean predictores útiles. Se dispone de varios programas de cómputo que llegan en secuencia a la denominada mejor ecuación de regresión, dependiendo de ciertos criterios. En la sección 12.9 profundizaremos en el estudio de esto.

Un criterio que suele utilizarse para ilustrar lo adecuado de un modelo ajustado de regresión es el **coeficiente de determinación múltiple** o $\boldsymbol{R^2}$.

Coeficiente de determinación múltiple o R^2

$$R^2 = \frac{SCR}{STCC} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SCE}{STCC}.$$

Advierta que esta descripción se parece a la que se hizo de R^2 en el capítulo 11. En este punto la explicación podría ser más clara, toda vez que ahora nos centramos en *SCR* como la **variabilidad explicada**. La cantidad R^2 tan sólo indica qué proporción de la variación total de la respuesta Y es explicada por el modelo ajustado. Con frecuencia los experimentadores reportan $R^2 \times 100\%$ e interpretan el resultado como el porcentaje de variación explicado con el modelo propuesto. La raíz cuadrada de R^2 se denomina **coeficiente de correlación múltiple** entre Y y el conjunto $x_1, x_2, ..., x_k$. En el ejemplo 12.4 el valor de R^2 que indica la proporción de variación explicada por las tres variables independientes x_1, x_2 y x_3 es

$$R^2 = \frac{SCR}{STCC} = \frac{399.45}{438.13} = 0.9117,$$

lo cual significa que 91.17% de la variación del porcentaje de supervivencia queda explicada por el modelo de regresión lineal.

La suma de cuadrados de regresión se puede emplear para obtener algún indicio acerca de si el modelo representa o no una explicación adecuada de la verdadera situación. Podemos probar la hipótesis H_0 de que la **regresión no es significativa** con sólo plantear la razón

$$f = \frac{SCR/k}{SCE/(n-k-1)} = \frac{SCR/k}{s^2}$$

y rechazar H_0 al nivel de significancia α cuando $f > f_\alpha(k, n-k-1)$. Para los datos del ejemplo 12.4 se obtiene

$$f = \frac{399.45/3}{4.298} = 30.98.$$

De la salida de resultados por computadora que aparece en la figura 12.1, el valor P es menor que 0.0001. Esto no debe malinterpretarse. Aunque indica que la regresión explicada por el modelo es significativa, no descarta la posibilidad de que

1. El modelo de regresión lineal en este conjunto de x no sea el único que se puede usar para explicar los datos; de hecho, quizás haya otros modelos con transformaciones sobre las x que generen un valor mayor para el estadístico F.
2. El modelo podría ser más eficaz si se incluyeran otras variables, además de x_1, x_2 y x_3, o quizá si se eliminaran una o más de las variables del modelo, por ejemplo x_3, que tiene un valor $P = 0.5916$.

El lector debería recordar el análisis de la sección 11.5 sobre las desventajas de utilizar R^2 como criterio para comparar modelos en competencia. Es evidente que dichas desventajas son relevantes en la regresión lineal múltiple. De hecho, los riesgos de su empleo en la regresión múltiple son aún mayores debido a que es muy grande la tentación de hacer un sobreajuste. Hay que tener siempre presente que $R^2 \approx 1.0$ siempre puede

obtenerse a expensas de los grados de libertad del error cuando se emplea un exceso de términos en el modelo. Sin embargo, $R^2 = 1$, que describe un modelo con ajuste casi perfecto, no siempre genera un modelo que hace buenas predicciones.

El coeficiente de determinación ajustado (R^2_{ajus})

En el capítulo 11 se presentan varias figuras que muestran listados de resultados por computadora, tanto del *SAS* como de *MINITAB*, en las que aparece un estadístico llamado R^2 *ajustado*, o un coeficiente de determinación ajustado. R^2 ajustado es una variación de R^2 que proporciona un **ajuste para los grados de libertad**. El coeficiente de determinación, según se definió en la página 407, no puede disminuir a medida que se agregan términos al modelo. En otras palabras, R^2 no disminuye a medida que se reducen los grados de libertad del error $n - k - 1$, ya que este último resultado se produce por un incremento de k, el número de términos en el modelo. R^2 ajustado se calcula dividiendo la *SCE* y la *STCC* entre sus grados de libertad respectivos de la siguiente manera.

R^2 ajustado

$$R^2_{\text{ajus}} = 1 - \frac{SCE\,/(n-k-1)}{STCC\,/(n-1)}.$$

Para ilustrar el uso de R^2_{ajus} se revisará el ejemplo 12.4.

¿Cómo la eliminación de x_3 afecta a R^2 y R^2_{ajus}?

La prueba t (o la prueba F correspondiente) para x_3 sugiere que un modelo más sencillo que sólo implique x_1 y x_2 bien podría ser una mejoría. En otras palabras, el modelo completo con todos los regresores podría estar sobreajustado. Por supuesto que es interesante investigar R^2 y R^2_{ajus} tanto para el modelo completo (x_1, x_2 y x_3) como para el modelo reducido (x_1, x_2). A partir de la figura 12.1 ya sabemos que $R^2_{\text{compl}} = 0.9117$. La *SCE* para el modelo reducido es 40.01, por lo que $R^2_{\text{reduc}} = 1 - \frac{40.01}{438.13} = 0.9087$. De esta forma, con x_3 dentro del modelo se explica más variabilidad. No obstante, como ya se dijo, esto ocurriría aun si el modelo estuviera sobreajustado. Desde luego que R^2_{ajus} está diseñada para proporcionar un estadístico que castigue un modelo sobreajustado, de manera que podríamos esperar que se favorezca al modelo restringido. Entonces, para el modelo completo

$$R^2_{\text{ajus}} = 1 - \frac{38.6764/9}{438.1308/12} = 1 - \frac{4.2974}{36.5109} = 0.8823,$$

mientras que para el modelo reducido (eliminación de x_3)

$$R^2_{\text{ajus}} = 1 - \frac{40.01/10}{438.1308/12} = 1 - \frac{4.001}{36.5109} = 0.8904.$$

Así, R^2_{ajus} realmente favorece el modelo reducido y confirma la evidencia proporcionada por las pruebas t y F, sugiriendo que el modelo reducido es preferible sobre el que contiene los tres regresores. El lector quizás espere que otros estadísticos sugieran el rechazo del modelo sobreajustado. Véase el ejercicio 12.40 de la página 471.

Prueba sobre un coeficiente individual

Agregar cualquier variable sencilla a un sistema de regresión *incrementará la suma de cuadrados de regresión* y con ello *se reducirá la suma de cuadrados del error*. En consecuencia, se debe decidir si el incremento en la regresión es suficiente para garantizar el uso de la variable en el modelo. Como es de esperarse, el empleo de variables sin importancia reduciría la eficacia de la ecuación de predicción incrementando la varianza de la respuesta estimada. Profundizaremos más en este punto al considerar la importancia de x_3 en el ejemplo 12.4. Inicialmente podemos probar

$$H_0: \beta_3 = 0,$$
$$H_1: \beta_3 \neq 0$$

usando la distribución t con 9 grados de libertad. Se tiene

$$t = \frac{b_3 - 0}{s\sqrt{c_{33}}} = \frac{-0.3433}{2.073\sqrt{0.0886}} = -0.556,$$

que indica que β_3 no difiere en forma significativa de cero y, por lo tanto, bien podríamos sentir que se justifica eliminar x_3 del modelo. Suponga que se considera la regresión de Y sobre el conjunto (x_1, x_2), las ecuaciones normales de mínimos cuadrados ahora se reducen a

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 \\ 59.43 & 394.7255 & 360.6621 \\ 81.82 & 360.6621 & 576.7264 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 377.50 \\ 1877.5670 \\ 2246.6610 \end{bmatrix}.$$

Los coeficientes de regresión estimados para este modelo reducido son

$$b_0 = 36.094, \quad b_1 = 1.031, \quad b_2 = -1.870,$$

y la suma de cuadrados de regresión resultante, con 2 grados de libertad, es

$$R(\beta_1, \beta_2) = 398.12.$$

Aquí se utiliza la notación $R(\beta_1, \beta_2)$ para indicar la suma de cuadrados de regresión del modelo restringido, y no debe confundirse con la *SCR*, es decir, la suma de cuadrados de regresión del modelo original con 3 grados de libertad. Entonces, la nueva suma de cuadrados del error es

$$STCC - R(\beta_1, \beta_2) = 438.13 - 398.12 = 40.01,$$

y el cuadrado medio del error resultante, con 10 grados de libertad, es

$$s^2 = \frac{40.01}{10} = 4.001.$$

¿Una prueba *t* de una variable tiene una prueba equivalente *F*?

En el ejemplo 12.4 la cantidad de variación en el porcentaje de supervivencia que se atribuye a x_3, en presencia de las variables x_1 y x_2, es

$$R(\beta_3 \mid \beta_1, \beta_2) = SCR - R(\beta_1, \beta_2) = 399.45 - 398.12 = 1.33,$$

que representa una pequeña proporción de toda la variación de la regresión. Esta cantidad de regresión agregada, como lo indica la prueba previa sobre β_3, es estadísticamente insignificante. Una prueba equivalente implica la formación de la razón

$$f = \frac{R(\beta_3 \mid \beta_1, \beta_2)}{s^2} = \frac{1.33}{4.298} = 0.309,$$

que es un valor de la distribución F con 1 y 9 grados de libertad. Recuerde que la relación básica entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad es

$$t^2 = f(1, v),$$

y se observa que el valor f de 0.309 es en realidad el cuadrado del valor t de -0.56.

Para generalizar los conceptos anteriores podemos evaluar el funcionamiento de una variable independiente x_i en el modelo general de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

observando la cantidad de regresión atribuida a x_i **sobre y por arriba de la atribuida a las demás variables**, es decir, la regresión sobre x_i *ajustada para las demás variables*. Por ejemplo, se dice que x_1 se evalúa calculando

$$R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k) = SCR - R(\beta_2, \beta_3, \ldots, \beta_k),$$

donde $R(\beta_2, \beta_3, \ldots, \beta_k)$ es la suma de cuadrados de regresión con $\beta_1 x_1$ eliminados del modelo. Para probar la hipótesis

$$H_0: \beta_1 = 0,$$
$$H_1: \beta_1 \neq 0,$$

se calcula

$$f = \frac{R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k)}{s^2},$$

y se compara con $f_\alpha(1, n - k - 1)$.

Pruebas *F* parciales en subconjuntos de coeficientes

De manera similar, se puede hacer una prueba para la significancia de un *conjunto* de las variables. Por ejemplo, para investigar simultáneamente la importancia de incluir x_1 y x_2 en el modelo se prueba la hipótesis

$$H_0: \beta_1 = \beta_2 = 0,$$
$$H_1: \beta_1 \text{ y } \beta_2 \text{ no son ambas cero},$$

calculando

$$f = \frac{[R(\beta_1, \beta_2 \mid \beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2} = \frac{[SCR - R(\beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2}$$

y comparando con $f_\alpha(2, n-k-1)$. El número de grados de libertad asociados con el numerador, en este caso 2, es igual al número de variables en el conjunto que se investiga.

Suponga que se desea probar la hipótesis

$$H_0: \beta_2 = \beta_3 = 0,$$
$$H_1: \beta_2 \text{ y } \beta_3 \text{ no son ambas cero}$$

para el ejemplo 12.4. Si desarrollamos el modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

podemos obtener $R(\beta_1) = SCR_{\text{reduc}} = 187.31179$. En la figura 12.1, de la página 459, tenemos $s^2 = 4.29738$ para el modelo completo. Por lo tanto, el valor de f para la prueba de hipótesis es

$$f = \frac{R(\beta_2, \beta_3 \mid \beta_1)/2}{s^2} = \frac{[R(\beta_1, \beta_2, \beta_3) - R(\beta_1)]/2}{s^2} = \frac{[\, SCR_{\text{compl}} - SCR_{\text{reduc}} \,]/2}{s^2}$$
$$= \frac{(399.45437 - 187.31179)/2}{4.29738} = 24.68278.$$

Esto implica que β_2 y β_3 no son iguales a cero de forma simultánea. Se puede utilizar un programa de estadística como el *SAS* para obtener el resultado anterior de manera directa, con un valor P de 0.0002. Los lectores deben observar que en los resultados de los programas de estadística para computadora aparecen valores P asociados con cada coeficiente individual del modelo. La hipótesis nula para cada una es que el coeficiente es igual a cero. Sin embargo, debemos señalar que la insignificancia de cualquier coeficiente no implica necesariamente que no deba ser incluido en el modelo final; sólo sugiere que es insignificante ante la presencia de todas las otras variables en el problema. El estudio de caso que se incluye al final del capítulo ilustra más esta cuestión.

12.7 Caso especial de ortogonalidad (opcional)

Antes de nuestro desarrollo original del problema general de regresión lineal se planteó la suposición de que las variables independientes se miden sin error y que con frecuencia están bajo el control del experimentador. A menudo ocurren como resultado de un *experimento diseñado con gran detalle*. De hecho, se puede incrementar la eficacia de la ecuación de predicción resultante utilizando un plan de experimentación adecuado.

Suponga que nuevamente consideramos la matriz **X,** tal como se definió en la sección 12.3. Podemos rescribirla como

$$\mathbf{X} = [\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}],$$

donde **1** representa una columna de unos y $\mathbf{x}_j$ es un vector columna que representa los niveles de x_j. Si

$$\mathbf{x}_p'\mathbf{x}_q = \mathbf{0}, \quad \text{para } p \neq q,$$

se dice que las variables x_p y x_q son *ortogonales* entre sí. Hay ciertas ventajas evidentes en tener una situación completamente ortogonal, en la cual $\mathbf{x}'_p\mathbf{x}_q = \mathbf{0}$.

para toda posible p y q, $p \neq q$ y, además,

$$\sum_{i=1}^{n} x_{ji} = 0, \quad j = 1, 2, \ldots, k.$$

La $\mathbf{X}'\mathbf{X}$ resultante es una matriz diagonal, y las ecuaciones normales de la sección 12.3 se reducen a

$$nb_0 = \sum_{i=1}^{n} y_i, \quad b_1 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{1i} y_i, \cdots, b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki} y_i.$$

Una ventaja importante es que es fácil hacer la partición de la *SCR* en **componentes de un solo grado de libertad**, cada uno de los cuales corresponde a la cantidad de variación de Y explicada por una variable controlada establecida. En la situación ortogonal se escribe

$$\begin{aligned} SCR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - b_0)^2 \\ &= b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2 \\ &= R(\beta_1) + R(\beta_2) + \cdots + R(\beta_k). \end{aligned}$$

La cantidad $R(\beta_i)$ es la cantidad de la suma de cuadrados de regresión asociada con un modelo que implica una sola variable independiente x_i.

Para probar simultáneamente la significancia de un conjunto de m variables en una situación ortogonal, la suma de cuadrados de regresión se convierte en

$$R(\beta_1, \beta_2, \ldots, \beta_m \mid \beta_{m+1}, \beta_{m+2}, \ldots, \beta_k) = R(\beta_1) + R(\beta_2) + \cdots + R(\beta_m),$$

y, por lo tanto,

$$R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k) = R(\beta_1)$$

se simplifica cuando se evalúa una sola variable independiente. Por consiguiente, la contribución de una variable determinada o un conjunto de variables se encuentra, en esencia, *ignorando* las demás variables del modelo. Las evaluaciones independientes del beneficio de las variables individuales se llevan a cabo usando las técnicas de análisis de varianza, tal como se presentan en la tabla 12.4. La variación total en la respuesta está dividida en componentes de un solo grado de libertad más el término del error con $n - k - 1$ grados de libertad. Cada valor f calculado se utiliza para probar una de las hipótesis

$$\left.\begin{array}{l} H_0: \ \beta_i = 0 \\ H_1: \ \beta_i \neq 0 \end{array}\right\} \qquad i = 1, 2, \ldots, k,$$

comparándolas con el punto crítico $f_\alpha(1, n - k - 1)$ o simplemente interpretando el valor P calculado a partir de la distribución f.

Tabla 12.4: Análisis de varianza para variables ortogonales

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
β_1	$R(\beta_1) = b_1^2 \sum_{i=1}^{n} x_{1i}^2$	1	$R(\beta_1)$	$\frac{R(\beta_1)}{s^2}$
β_2	$R(\beta_2) = b_2^2 \sum_{i=1}^{n} x_{2i}^2$	1	$R(\beta_2)$	$\frac{R(\beta_2)}{s^2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_k	$R(\beta_k) = b_k^2 \sum_{i=1}^{n} x_{ki}^2$	1	$R(\beta_k)$	$\frac{R(\beta_k)}{s^2}$
Error	SCE	$n-k-1$	$s^2 = \frac{SCE}{n-k-1}$	
Total	$SST = S_{yy}$	$n-1$		

Ejemplo 12.8: Suponga que un científico recaba datos experimentales sobre el radio de un grano propulsor, Y, en función de la temperatura del polvo, x_1, la tasa de extrusión, x_2, y la temperatura del molde, x_3. Ajuste un modelo de regresión lineal para predecir el radio del grano y determine la eficacia de cada variable que interviene en el modelo. Los datos se presentan en la tabla 12.5.

Tabla 12.5: Datos para el ejemplo 12.8

Radio del grano	Temperatura del polvo	Tasa de extrusión	Temperatura del molde
82	150 (−1)	12 (−1)	220 (−1)
93	190 (+1)	12 (−1)	220 (−1)
114	150 (−1)	24 (+1)	220 (−1)
124	150 (−1)	12 (−1)	250 (+1)
111	190 (+1)	24 (+1)	220 (−1)
129	190 (+1)	12 (−1)	250 (+1)
157	150 (−1)	24 (+1)	250 (+1)
164	190 (+1)	24 (+1)	250 (+1)

***Solución*:** Observe que cada variable está controlada en dos niveles, y que el experimento está compuesto por las ocho combinaciones posibles. Por conveniencia, los datos de las variables independientes se codificaron mediante las siguientes fórmulas:

$$x_1 = \frac{\text{temperatura del polvo} - 170}{20},$$
$$x_2 = \frac{\text{tasa de extrusión} - 18}{6},$$
$$x_3 = \frac{\text{temperatura del molde} - 235}{15}.$$

Los niveles resultantes de x_1, x_2 y x_3 toman los valores -1 y $+1$, tal como se indica en la tabla con los datos. Este diseño experimental en particular permite la ortogonalidad que

queremos ilustrar aquí. (En el capítulo 15 se analiza un tratamiento más completo de este tipo de diseño experimental). La matriz **X** es

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

y las condiciones de ortogonalidad se verifican con facilidad.

Ahora podemos calcular los coeficientes

$$b_0 = \frac{1}{8}\sum_{i=1}^{8} y_i = 121.75, \qquad b_1 = \frac{1}{8}\sum_{i=1}^{8} x_{1i}y_i = \frac{20}{8} = 2.5,$$

$$b_2 = \frac{\sum_{i=1}^{8} x_{2i}y_i}{8} = \frac{118}{8} = 14.75, \qquad b_3 = \frac{\sum_{i=1}^{8} x_{3i}y_i}{8} = \frac{174}{8} = 21.75,$$

de manera que, en términos de las variables codificadas, la ecuación de predicción es

$$\hat{y} = 121.75 + 2.5\,x_1 + 14.75\,x_2 + 21.75\,x_3.$$

El análisis de varianza de la tabla 12.6 presenta las contribuciones independientes a la *SCR* de cada variable. Cuando los resultados se comparan con $f_{0.05}(1.4)$, cuyo valor es 7.71, indican que x_1 no contribuye de manera significativa a un nivel de 0.05; mientras que las variables x_2 y x_3 sí son significativas. En este ejemplo el estimado para σ^2 es 23.1250. Igual que en el caso de una sola variable independiente, se debe señalar que este estimado no sólo contiene variación por el error experimental, a menos que el modelo postulado sea correcto. De otra manera, el estimado estará "contaminado" por la falta de ajuste, además del error puro, y la falta de ajuste sólo se puede separar si se obtienen múltiples observaciones experimentales para las distintas combinaciones (x_1, x_2, x_3).

Tabla 12.6: Análisis de varianza para los datos del radio de los granos

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrados medios**	***f* calculada**	**Valor *P***
β_1	$(2.5)^2(8) = 50.00$	1	50.00	2.16	0.2156
β_2	$(14.75)^2(8) = 1740.50$	1	1740.50	75.26	0.0010
β_3	$(21.75)^2(8) = 3784.50$	1	3784.50	163.65	0.0002
Error	92.50	4	23.13		
Total	5667.50	7			

Como x_1 no es significativa, simplemente se puede eliminar del modelo sin alterar los efectos de las otras variables. Observe que tanto x_2 como x_3 tienen un efecto positivo sobre el radio del grano, pero x_3 es el factor más importante debido a la pequeñez de su valor *P*.

Ejercicios

12.31 Calcule e interprete el coeficiente de determinación múltiple para las variables del ejercicio 12.1 de la página 450.

12.32 Pruebe si la regresión explicada por el modelo del ejercicio 12.1, que se encuentra en la página 450, es significativa a un nivel de significancia de 0.01.

12.33 Pruebe si la regresión explicada por el modelo del ejercicio 12.5, de la página 450, es significativa a un nivel de significancia de 0.01.

12.34 Para el modelo del ejercicio 12.5 de la página 450 pruebe la hipótesis

$$H_0\colon\ \beta_1 = \beta_2 = 0,$$
$$H_1\colon\ \beta_1 \text{ y } \beta_2 \text{ no son ambas cero.}$$

12.35 Repita el ejercicio 12.17 de la página 461 usando el estadístico *F*.

12.36 Se realizó un pequeño experimento para ajustar una ecuación de regresión múltiple que relaciona el producto, *y,* con la temperatura, x_1, el tiempo de reacción, x_2, y la concentración de uno de los reactantes, x_3. Se eligieron dos niveles de cada variable y se registraron las siguientes mediciones correspondientes a las variables independientes codificadas:

y	x_1	x_2	x_3
7.6	−1	−1	−1
8.4	1	−1	−1
9.2	−1	1	−1
10.3	−1	−1	1
9.8	1	1	−1
11.1	1	−1	1
10.2	−1	1	1
12.6	1	1	1

a) Utilice las variables codificadas para estimar la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

b) Divida la *SCR*, es decir, la suma de cuadrados de regresión, en tres componentes de un solo grado de libertad atribuibles a x_1, x_2 y x_3, respectivamente. Construya una tabla de análisis de varianza que indique las pruebas de significancia para cada variable.

12.37 Considere los datos de energía eléctrica del ejercicio 12.5 de la página 450. Pruebe H_0: $\beta_1 = \beta_2 = 0$ utilizando $R(\beta_1, \beta_2 \mid \beta_3, \beta_4)$. Proporcione un valor *P* y saque conclusiones.

12.38 Considere los datos del ejercicio 12.36. Calcule lo siguiente:

$$R(\beta_1 \mid \beta_0), \qquad R(\beta_1 \mid \beta_0, \beta_2, \beta_3),$$
$$R(\beta_2 \mid \beta_0, \beta_1), \qquad R(\beta_2 \mid \beta_0, \beta_1, \beta_3),$$
$$R(\beta_3 \mid \beta_0, \beta_1, \beta_2), \quad R(\beta_1, \beta_2 \mid \beta_3).$$

Haga comentarios al respecto.

12.39 Considere los datos del ejercicio 11.55 de la página 437. Ajuste un modelo de regresión utilizando el peso y el cociente de manejo como variables explicativas. Compare este modelo con el de la RLS (regresión lineal simple) utilizando sólo el peso. Utilice R^2, R^2_{ajus} y cualquier estadístico *t* (o *F*) que necesite para comparar la RLS con el modelo de regresión múltiple.

12.40 Considere el ejemplo 12.4. La figura 12.1 de la página 459 presenta una salida de resultados del *SAS* para un análisis del modelo que contiene las variables x_1, x_2 y x_3. Céntrese en el intervalo de confianza de la respuesta media μ_Y en las ubicaciones (x_1, x_2, x_3) que representan los 13 puntos de los datos. Considere el elemento en la salida de resultados indicado con C.V., que representa al **coeficiente de variación**, el cual se define como

$$\text{C.V.} = \frac{s}{\bar{y}} \cdot 100,$$

donde $s = \sqrt{s^2}$ es la **raíz del cuadrado medio del error**. El coeficiente de variación se utiliza con frecuencia como otro criterio para comparar modelos en competencia. Se trata de una cantidad sin escala que expresa al estimado de σ, es decir, *s*, como un porcentaje de la respuesta promedio $\bar{y}$. Al competir por el "mejor" modelo de un grupo de modelos en competencia se busca un modelo con un valor pequeño de C.V. Haga un análisis de regresión del conjunto de datos que se presenta en el ejemplo 12.4, pero elimine x_3. Compare el modelo completo (x_1, x_2, x_3) con el restringido (x_1, x_2) y céntrese en dos criterios: i) C.V.; ii) la anchura de los intervalos de confianza sobre μ_Y. Para el segundo criterio usted quizá desearía usar la anchura promedio. Haga comentarios al respecto.

12.41 Considere el ejemplo 12.3 de la página 447. Compare los dos modelos en competencia

$$\text{Primer orden:}\quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,$$
$$\text{Segundo orden:}\quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i.$$

Utilice R^2_{ajus} para realizar la comparación. Pruebe H_0: $\beta_{11} = \beta_{22} = \beta_{12} = 0$. También utilice el C.V. que se mencionó en el ejercicio 12.40.

12.42 En el ejemplo 12.8 se trata el caso de eliminar del modelo x_1, que representa la temperatura del polvo, ya que el valor P basado en la prueba F es 0.2156, en tanto que los valores P para x_2 y x_3 son casi cero.

a) Reduzca el modelo eliminando x_1, después genere un modelo completo y uno restringido (o reducido), y compárelos basándose en R^2_{ajus}.

b) Compare los modelos completo y restringido usando intervalos de predicción de 95% de ancho sobre una nueva observación. El "mejor" de ambos modelos será aquel con intervalos de predicción más "estrechos". Utilice el promedio del ancho de los intervalos de predicción.

12.43 Considere los datos del ejercicio 12.13 de la página 452. ¿La respuesta, o sea el uso, se puede explicar en forma adecuada mediante una sola variable (ya sea la viscosidad o la carga) con una *RLS* en vez de con la regresión completa con dos variables? Justifique su respuesta con pruebas de hipótesis, así como con la comparación de los tres modelos en competencia.

12.44 Para el conjunto de datos que se da en el ejercicio 12.16 de la página 453, ¿es posible explicar la respuesta en forma adecuada usando dos variables regresoras cualesquiera? Analice el problema.

12.8 Variables categóricas o indicadoras

Un caso especial de aplicación muy importante de la regresión lineal múltiple ocurre cuando una o más de las variables regresoras son **variables categóricas, indicadoras** o **ficticias**. Es probable que en un proceso químico el ingeniero desee modelar el producto del proceso en comparación con regresores tales como la temperatura del proceso y el tiempo de reacción. Sin embargo, hay interés por el uso de dos catalizadores diferentes y por incluir de algún modo el "catalizador" en el modelo. El efecto del catalizador no se puede medir sobre un continuo, de manera que es una variable categórica. Un analista podría desear modelar el precio de casas en comparación con regresores que incluyan los pies cuadrados de superficie habitable, x_1, la superficie del terreno, x_2, y la antigüedad de la vivienda, x_3. Estos regresores son de naturaleza claramente continua. Sin embargo, es evidente que el costo de las casas podría variar en forma sustancial de una zona del país a otra. Si reuniéramos datos acerca de casas en el este, el medio oeste, en el sur y en el oeste, tendríamos una variable indicadora con **cuatro categorías**. En el ejemplo del proceso químico, si utilizáramos dos catalizadores tendríamos una variable indicadora con dos categorías. En un ejemplo biomédico, donde se compara un medicamento con un placebo, a todos los sujetos se les evalúa con varias mediciones continuas, como su edad, presión sanguínea, etcétera, al igual que el género, que por supuesto es una variable categórica con dos categorías. De esta manera, además de las variables continuas existen dos variables indicadoras, el tratamiento con dos categorías (medicamento activo y placebo) y el género con dos categorías (hombre y mujer).

Modelo con variables categóricas

Para ilustrar la forma en que las variables indicadoras participan en el modelo utilizaremos el ejemplo del proceso químico. Suponga que y = producto, x_1 = temperatura y x_2 = tiempo de reacción. Ahora denotaremos con z la variable indicadora. Sea $z = 0$ para el catalizador 1 y $z = 1$ para el catalizador 2. La asignación del indicador (0, 1) al catalizador es arbitraria. Como resultado, el modelo se convierte en

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 z_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Tres categorías

Continuamos aplicando la estimación de los coeficientes con el método de los mínimos cuadrados. En el caso de tener tres niveles o categorías de una sola variable indicadora,

el modelo incluirá **dos** regresores, digamos z_1 y z_2, donde la asignación (0, 1) es como sigue:

$$\begin{bmatrix} z_1 & z_2 \\ \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

donde **0** y **1** son vectores de ceros y unos, respectivamente. En otras palabras, si hay ℓ categorías, el modelo incluye $\ell - 1$ términos reales.

Puede ser aleccionador observar la representación gráfica del modelo con 3 categorías. En aras de la simplicidad, se considerará una sola variable continua x. Como resultado, el modelo quedará representado como

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i.$$

Así, la figura 12.2 refleja la naturaleza del modelo. Las siguientes son expresiones del modelo para las tres categorías.

$$\begin{aligned} E(Y) &= (\beta_0 + \beta_2) + \beta_1 x, && \text{categoría 1,} \\ E(Y) &= (\beta_0 + \beta_3) + \beta_1 x, && \text{categoría 2,} \\ E(Y) &= \beta_0 + \beta_1 x, && \text{categoría 3.} \end{aligned}$$

Como resultado, el modelo que incluye variables categóricas en esencia implica un **cambio en la intersección** a medida que se pasa de una categoría a otra. Desde luego, aquí se asume que los **coeficientes de las variables continuas son los mismos entre las categorías**.

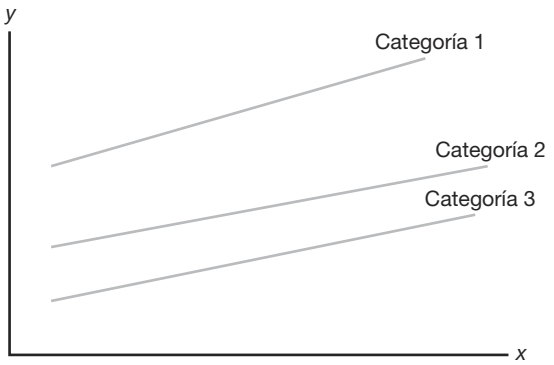


Figura 12.2: Caso de tres categorías.

Ejemplo 12.9: Considere los datos de la tabla 12.7. La respuesta y es la cantidad de sólidos en suspensión en un sistema de limpieza de carbón. La variable x es el pH del sistema y se utilizan tres polímeros diferentes. Así, "polímero" es categórico con tres categorías, de manera que produce dos términos en el modelo, el cual queda como

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i, \qquad i = 1, 2, \ldots, 18.$$

Luego, tenemos

$$z_1 = \begin{cases} 1, & \text{para el polímero 1,} \\ 0, & \text{en cualquier caso,} \end{cases} \quad \text{y} \quad z_2 = \begin{cases} 1, & \text{para el polímero 2,} \\ 0, & \text{en cualquier caso.} \end{cases}$$

A partir del análisis de la figura 12.3 se sacan las siguientes conclusiones. El coeficiente b_1 para el pH es el estimado de la **pendiente común** que se asume en el análisis de regresión. Todos los términos del modelo son estadísticamente significativos. Así, el pH y la naturaleza del polímero tienen un efecto sobre la cantidad de limpieza. Los signos y las magnitudes de los coeficientes de z_1 y z_2 indican que el polímero más eficaz para la limpieza es el polímero 1 (produce más sólidos en suspensión), seguido por el polímero 2, y que el menos eficaz es el polímero 3.

Tabla 12.7: Datos para el ejemplo 12.9

x (pH)	y (cantidad de sólidos en suspensión)	Polímero
6.5	292	1
6.9	329	1
7.8	352	1
8.4	378	1
8.8	392	1
9.2	410	1
6.7	198	2
6.9	227	2
7.5	277	2
7.9	297	2
8.7	364	2
9.2	375	2
6.5	167	3
7.0	225	3
7.2	247	3
7.6	268	3
8.7	288	3
9.2	342	3

La pendiente puede variar con las categorías indicadoras

En el análisis efectuado hasta el momento se ha supuesto que los términos de las variables indicadoras entran al modelo en forma aditiva, lo cual sugiere que las pendientes, como las que se aprecian en la figura 12.2, son constantes en todas las categorías. Es evidente que éste no siempre será el caso. Existe la posibilidad de que las pendientes varíen y realmente se ponga a prueba esta condición de **paralelismo** al incluir términos de producto o **interacción** entre los términos indicadores y las variables continuas. Por ejemplo, suponga que se eligen un modelo con un regresor continuo y una variable indicadora con dos niveles. El modelo entonces quedaría como sigue

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon.$$

```
                          Sum of
          Source   DF       Squares   Mean Square     F Value     Pr > F
          Mode l    3    80181.73127   26727.24376       73.68     <.0001
           Error   14     5078.71318     362.76523
Corrected  Total   17    85260.44444

 R-Square      Coeff  Var       Root  MSE          y Mean
 0.940433       6.316049       19.04640        301.5556

                                                  Standard
Parameter       Estimate          Error   t Value   Pr >  |t|
Intercept  -161.8973333    37.43315576     -4.32      0.0007
        x    54.2940260     4.75541126     11.42      <.0001
       z1    89.9980606    11.05228237      8.14      <.0001
       z2    27.1656970    11.01042883      2.47      0.0271
```

Figura 12.3: Salida de resultados del *SAS* para el ejemplo 12.9.

Este modelo sugiere que para la categoría 1 ($z = 1$),

$$E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x,$$

mientras que para la categoría 2 ($z = 0$),

$$E(y) = \beta_0 + \beta_1 x.$$

Por consiguiente, se permite que varíen la intersección y las pendientes para las dos categorías. En la figura 12.4 se presentan las rectas de regresión con pendientes variables para las dos categorías.

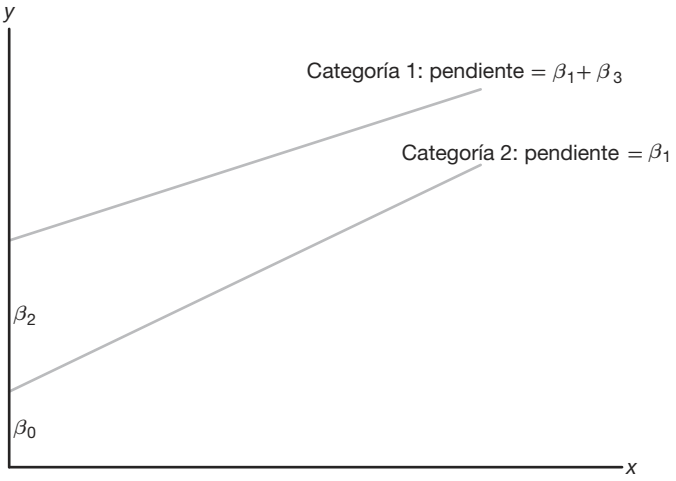


Figura 12.4: Falta de paralelismo en las variables categóricas.

En este caso β_0, β_1 y β_2 son positivas, mientras que β_3 es negativa con $|\beta_3| < \beta_1$. Por supuesto, si el coeficiente de interacción β_3 es insignificante, regresamos al modelo común de la pendiente.

Ejercicios

12.45 Se realizó un estudio para evaluar el combustible que se ahorra al conducir un automóvil sedán de cuatro puertas en vez de una camioneta o un SUV (vehículo deportivo utilitario). Las variables continuas son la lectura del odómetro y el octanaje de la gasolina empleada. La variable de respuesta se da en millas por galón. Los datos se presentan a continuación.

MP	Tipo de automóvil	Odómetro	Octanaje
34.5	sedan	75,000	87.5
33.3	sedan	60,000	87.5
30.4	sedan	88,000	78.0
32.8	sedan	15,000	78.0
35.0	sedan	25,000	90.0
29.0	sedan	35,000	78.0
32.5	sedan	102,000	90.0
29.6	sedan	98,000	87.5
16.8	van	56,000	87.5
19.2	van	72,000	90.0
22.6	van	14,500	87.5
24.4	van	22,000	90.0
20.7	van	66,500	78.0
25.1	van	35,000	90.0
18.8	van	97,500	87.5
15.8	van	65,500	78.0
17.4	van	42,000	78.0
15.6	SUV	65,000	78.0
17.3	SUV	55,500	87.5
20.8	SUV	26,500	87.5
22.2	SUV	11,500	90.0
16.5	SUV	38,000	78.0
21.3	SUV	77,500	90.0
20.7	SUV	19,500	78.0
24.1	SUV	87,000	90.0

a) Ajuste un modelo de regresión lineal que incluya dos variables indicadoras. Utilice (0, 0) para denotar al sedán de cuatro puertas.

b) ¿Qué tipo de vehículo parece tener un mayor rendimiento del combustible?

c) Analice la diferencia entre una camioneta y un SUV en términos del rendimiento del combustible.

12.46 Se efectuó un estudio para determinar si el género del titular de la tarjeta de crédito era un factor importante en la generación de utilidades para cierta empresa de tarjetas de crédito. Las variables consideradas fueron el ingreso, el número de miembros de la familia y el género del titular de la tarjeta. Los datos son los siguientes:

Utilidad	Ingreso	Género	Miembros de la familia
157	45,000	M	1
−181	55,000	M	2
−253	45,800	M	4
158	38,000	M	3
75	75,000	M	4
202	99,750	M	4
−451	28,000	M	1
146	39,000	M	2
89	54,350	M	1
−357	32,500	M	1
522	36,750	F	1
78	42,500	F	3
5	34,250	F	2
−177	36,750	F	3
123	24,500	F	2
251	27,500	F	1
−56	18,000	F	1
453	24,500	F	1
288	88,750	F	1
−104	19,750	F	2

a) Ajuste un modelo de regresión lineal usando las variables disponibles. Con base en el modelo ajustado, ¿la empresa preferiría clientes del género masculino o del femenino?

b) ¿Diría usted que el ingreso fue un factor importante para explicar la variabilidad de la utilidad?

12.9 Métodos secuenciales para la selección del modelo

En ocasiones las pruebas de significancia estudiadas en la sección 12.6 son muy adecuadas para determinar cuáles variables se deben usar en el modelo final de regresión. Dichas pruebas sin duda son eficaces si el experimento se puede planear y las variables son ortogonales entre sí. Incluso si las variables no son ortogonales, las pruebas *t* individuales se pueden usar en muchos problemas en donde se investigan pocas variables. Sin embargo, existen muchos problemas en los que es necesario utilizar técnicas más elaboradas para seleccionar las variables, en particular si el experimento exhibe una desviación sustancial de la ortogonalidad. Los coeficientes de correlación de la muestra $r_{x_i x_j}$ proporcionan medidas útiles de **multicolinealidad** (dependencia lineal) entre las

variables independientes. Como sólo estamos interesados en la dependencia lineal entre variables independientes, no nos confundiremos si eliminamos las x de la notación y sólo escribimos $r_{x_i x_j} = r_{ij}$, donde

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}.$$

Observe que, en sentido estricto, las r_{ij} no proporcionan estimados verdaderos de los coeficientes de correlación de la población, ya que las x en realidad no son variables aleatorias en el contexto que se estudia aquí. Así, el término *correlación*, aunque estándar, quizá sea inadecuado.

Cuando uno o más de esos coeficientes de correlación muestral se desvía de manera sustancial de cero, suele ser muy difícil encontrar el subconjunto de variables más eficaz para incluirlo en la ecuación de predicción. De hecho, en ciertos problemas la multicolinealidad es tan extrema que no es posible encontrar un predictor adecuado, a menos que se investiguen todos los subconjuntos posibles de variables. En la bibliografía se citan los análisis informativos de Hocking (1976) para la selección de modelos de regresión. En la obra de Myers (1990), también citado, se estudian procedimientos para detectar la multicolinealidad.

El usuario de la regresión lineal múltiple busca lograr uno de tres objetivos:

1. Obtener estimados de coeficientes individuales en un modelo completo.
2. Estudiar variables para determinar cuáles tienen un efecto significativo sobre la respuesta.
3. Calcular la ecuación de predicción más eficaz.

En 1) se sabe de antemano que todas las variables deben incluirse en el modelo. En 2) la predicción es secundaria; mientras que en 3) los coeficientes de regresión individuales no son tan importantes como la calidad de la respuesta estimada $\hat{y}$. Para cada una de las situaciones anteriores la multicolinealidad en el experimento puede tener un efecto profundo sobre el éxito de la regresión.

En esta sección se estudian algunos procedimientos secuenciales estándar para seleccionar variables, los cuales se basan en la idea de que una sola variable o un conjunto de ellas no debería aparecer en la ecuación de estimación, a menos que origine un incremento significativo en la suma de cuadrados de regresión o, en forma equivalente, un incremento significativo de R^2, el coeficiente de determinación múltiple.

Ilustración de la selección de las variables en presencia de colinealidad

Ejemplo 12.10: Considere los datos de la tabla 12.8, que muestra mediciones de 9 bebés. El objetivo del experimento era calcular una ecuación de estimación apropiada que relacionara la talla del bebé con todas las variables independientes o un subconjunto de ellas. Los coeficientes de correlación muestral, que indican la dependencia lineal entre las variables independientes, se incluyen en la matriz simétrica

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 \end{matrix} \\ \begin{bmatrix} 1.0000 & 0.9523 & 0.5340 & 0.3900 \\ 0.9523 & 1.0000 & 0.2626 & 0.1549 \\ 0.5340 & 0.2626 & 1.0000 & 0.7847 \\ 0.3900 & 0.1549 & 0.7847 & 1.0000 \end{bmatrix} \end{array}$$

Tabla 12.8: Datos relacionados con la talla de bebés*

Talla del bebé, y (cm)	Edad, x_1 (días)	Talla al nacer, x_2 (cm)	Peso al nacer, x_3 (kg)	Tamaño del pecho al nacer, x_4 (cm)
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67.0	88	49.0	5.52	36.5
53.5	67	43.0	3.21	27.2
62.7	80	48.0	4.32	27.7
56.2	74	48.0	2.31	28.3
68.5	94	53.0	4.30	30.3
69.2	102	58.0	3.71	28.7

*Datos analizados por el Statistical Consulting Center, Virginia Tech, Blacksburg, Virginia.

Observe que parece haber una cantidad apreciable de multicolinealidad. Se utilizó la técnica de mínimos cuadrados descrita en la sección 12.2 y se usó el modelo completo para ajustar la ecuación de regresión estimada, la cual quedó como sigue:

$$\hat{y} = 7.1475 + 0.1000x_1 + 0.7264x_2 + 3.0758x_3 - 0.0300x_4.$$

El valor de s^2 con 4 grados de libertad es 0.7414, y se encontró que el valor del coeficiente de determinación para este modelo es 0.9908. En la tabla 12.9 se proporciona la suma de cuadrados de regresión que mide la variación atribuida a cada variable individual en presencia de las demás, así como los valores t correspondientes.

Tabla 12.9: Valores t para los datos de regresión de la tabla 12.8

Variable x_1	Variable x_2	Variable x_3	Variable x_4
$R(\beta_1 \mid \beta_2, \beta_3, \beta_4)$	$R(\beta_2 \mid \beta_1, \beta_3, \beta_4)$	$R(\beta_3 \mid \beta_1, \beta_2, \beta_4)$	$R(\beta_4 \mid \beta_1, \beta_2, \beta_3)$
$= 0.0644$	$= 0.6334$	$= 6.2523$	$= 0.0241$
$t = 0.2947$	$t = 0.9243$	$t = 2.9040$	$t = -0.1805$

Una región crítica de dos colas, con 4 grados de libertad y un nivel de significancia de 0.05, es dada por $|t| > 2.776$. De los cuatro valores t calculados **sólo la variable x_3 parece ser significativa**. Sin embargo, recuerde que aunque el estadístico t descrito en la sección 12.6 mide el beneficio que aporta una variable ajustada a todas las demás, no detecta la importancia potencial de una variable en combinación con un subconjunto de variables. Por ejemplo, considere el modelo sólo con las variables x_2 y x_3 en la ecuación. El análisis de los datos proporciona la función de regresión

$$\hat{y} = 2.1833 + 0.9576x_2 + 3.3253x_3,$$

con $R^2 = 0.9905$, que por supuesto no es una reducción sustancial de $R^2 = 0.9907$ para el modelo completo. Sin embargo, a menos que las características del desempeño de esta combinación particular hayan sido observadas, no estaríamos conscientes de su potencial predictivo. Esto, desde luego, apoya una metodología que observe *todas las regresiones posibles*, o un procedimiento secuencial sistemático diseñado para probar subconjuntos diferentes.

Regresión por etapas

Un procedimiento estándar para buscar el "subconjunto óptimo" de variables ante la ausencia de ortogonalidad es una técnica denominada **regresión por etapas**, que se basa en el procedimiento de introducir en forma secuencial las variables al modelo, una por una. Dado un tamaño α predeterminado, la descripción de la rutina por etapas se entenderá mejor si primero se describen los métodos de **selección hacia delante** y **eliminación hacia atrás**.

La **selección hacia delante** se basa en el concepto de que las variables deben insertarse una por una hasta obtener una ecuación de regresión satisfactoria. El procedimiento es como sigue:

PASO 1. Elija la variable que proporcione la mayor suma de cuadrados de regresión cuando se ejecute la regresión lineal simple con y o, en forma equivalente, aquella que proporcione el mayor valor de R^2. Esta variable inicial se llamará x_1. Si x_1 es insignificante, el procedimiento se suspende.

PASO 2. Seleccione la variable que al ser integrada al modelo proporciona el mayor incremento de R^2, en presencia de x_1, sobre la R^2 encontrada en el paso 1. Ésta, por supuesto, es la variable x_j para la que

$$R(\beta_j|\beta_1) = R(\beta_1, \beta_j) - R(\beta_1)$$

es más grande. Dicha variable se llamará x_2. Luego se ajusta el modelo de regresión con x_1 y x_2, y se observa R^2. Si x_2 es insignificante, el procedimiento se suspende.

PASO 3. Elija la variable x_j que proporciona el valor más grande de

$$R(\beta_j \mid \beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_j) - R(\beta_1, \beta_2),$$

otra vez da como resultado el incremento mayor de R^2 sobre el que se obtuvo en el paso 2. A esta variable se le denomina x_3, y ahora se tiene un modelo de regresión que incluye x_1, x_2 y x_3. Si x_3 es insignificante, el procedimiento se suspende.

Este proceso continúa hasta que la variable más reciente incluida ya no produce un incremento significativo en la regresión explicada. Tal incremento se puede determinar en cada paso utilizando adecuadamente una prueba F o una prueba t parciales. Por ejemplo, en el paso 2 el valor

$$f = \frac{R(\beta_2|\beta_1)}{s^2}$$

se determina para probar la pertinencia de x_2 en el modelo. Aquí, el valor de s^2 es el cuadrado medio del error para el modelo que contiene las variables x_1 y x_2. De manera similar, en el paso 3 la razón

$$f = \frac{R(\beta_3 \mid \beta_1, \beta_2)}{s^2}$$

prueba la pertinencia de x_3 en el modelo. Sin embargo, ahora el valor de s^2 es el cuadrado medio del error para el modelo que contiene las tres variables x_1, x_2 y x_3. Si en el paso 2, $f < f_\alpha(1, n-3)$ para un nivel de significancia preseleccionado, x_2 no está incluida y el

proceso finaliza, lo que da como resultado una ecuación lineal simple que relaciona y y x_1. Sin embargo, si $f > f_\alpha(1, n-3)$, se avanza al paso 3. De nuevo, si en el paso 3, $f < f_\alpha(1, n-4)$, entonces x_3 no se incluye y el proceso termina con la ecuación de la regresión apropiada que contiene las variables x_1 y x_2.

La **eliminación hacia atrás** implica los mismos conceptos que la selección hacia delante, excepto que se comienza con todas las variables en el modelo. Por ejemplo, suponga que hay cinco variables en consideración. Los pasos son:

PASO 1. Ajuste una ecuación de regresión con las cinco variables incluidas en el modelo. Elija la variable que proporcione el valor más pequeño de la suma de cuadrados de regresión **ajustada para las demás**. Suponga que dicha variable es x_2. Elimine x_2 del modelo si

$$f = \frac{R(\beta_2 \mid \beta_1, \beta_3, \beta_4, \beta_5)}{s^2}$$

es insignificante.

PASO 2. Ajuste una ecuación de regresión utilizando las variables restantes x_1, x_3, x_4 y x_5, y repita el paso 1. Suponga que esta vez elige la variable x_5. Nuevamente, si

$$f = \frac{R(\beta_5 \mid \beta_1, \beta_3, \beta_4)}{s^2}$$

es insignificante, se retira del modelo la variable x_5. En cada paso la s^2 que se usa en la prueba F es el cuadrado medio del error para el modelo de regresión en esa etapa.

Este proceso se repite hasta que en algún paso la variable con la suma de cuadrados de regresión ajustada más pequeña produce un valor f significativo a un nivel de significancia predeterminado.

La **regresión por etapas** se lleva a cabo con una modificación ligera pero importante del procedimiento de selección hacia delante. La modificación requiere efectuar más pruebas en cada etapa para garantizar la eficacia continuada de las variables que se hubieran incluido en el modelo durante alguna etapa anterior. Esto representa una mejoría sobre la selección hacia delante, ya que es muy posible que una variable que haya entrado a la ecuación de regresión en una etapa temprana resulte poco importante o redundante debido a las relaciones que existen entre ella y las otras variables que se incluyeron en etapas posteriores. Por lo tanto, en la etapa en que se incluyó una variable nueva a la ecuación de regresión mediante un incremento significativo de R^2, según lo determina la prueba F, todas las variables que ya estén en el modelo se someten a pruebas F (o bien, a pruebas t) a la luz de esta nueva variable, y si no muestran un valor f significativo, se eliminan. El procedimiento continúa hasta que se alcance una etapa donde ya no sea posible insertar ni eliminar variables adicionales. Este procedimiento por etapas se ilustra con el siguiente ejemplo.

Ejemplo 12.11: Utilice las técnicas de regresión por etapas y calcule un modelo de regresión lineal adecuado para predecir la talla de los bebés cuyos datos se presentan en la tabla 12.8.

***Solución*:** **PASO 1.** Se considera cada variable por separado y se ajustan cuatro ecuaciones individuales de regresión lineal simple. Se calculan las siguientes sumas de cua-

drados de regresión pertinentes:

$$R(\beta_1) = 288.1468, \qquad R(\beta_2) = 215.3013,$$
$$R(\beta_3) = 186.1065, \qquad R(\beta_4) = 100.8594.$$

Es evidente que la variable x_1 proporciona la suma de cuadrados de regresión más elevada. El cuadrado medio del error para la ecuación que implica sólo x_1 es $s^2 = 4.7276$, y como

$$f = \frac{R(\beta_1)}{s^2} = \frac{288.1468}{4.7276} = 60.9500,$$

que excede a $f_{0.05}(1,7) = 5.59$, la variable x_1 es significativa y se introduce al modelo.

PASO 2. En esta etapa se ajustan tres ecuaciones de regresión y todas incluyen a x_1. Los resultados importantes para las combinaciones (x_1, x_2), (x_1, x_3) y (x_1, x_4) son

$$R(\beta_2|\beta_1) = 23.8703, \quad R(\beta_3|\beta_1) = 29.3086, \quad R(\beta_4|\beta_1) = 13.8178.$$

La variable x_3 muestra la mayor suma de cuadrados de regresión en presencia de x_1. La regresión que implica x_1 y x_3 proporciona un valor nuevo de $s^2 = 0.6307$, y como

$$f = \frac{R(\beta_3|\beta_1)}{s^2} = \frac{29.3086}{0.6307} = 46.47,$$

que excede a $f_{0.05}(1, 6) = 5.99$, la variable x_3 es significativa y se incluye en el modelo junto con x_1. Ahora debemos someter a x_1 a una prueba de significancia en presencia de x_3. Encontramos que $R(\beta_1 \mid \beta_3) = 131.349$, en consecuencia,

$$f = \frac{R(\beta_1|\beta_3)}{s^2} = \frac{131.349}{0.6307} = 208.26,$$

que es muy significativa. Por lo tanto, se mantiene x_1 junto con x_3.

PASO 3. Con x_1 y x_3 incluidas en el modelo, ahora se requiere $R(\beta_2 \mid \beta_1, \beta_3)$ y $R(\beta_4 \mid \beta_1, \beta_3)$ para determinar cuál de las dos variables restantes, si es que acaso se puede incluir alguna, se debe incluir en esta etapa. Del análisis de regresión, usando x_2 junto con x_1 y x_3, se observa que $R(\beta_2 \mid \beta_1, \beta_3) = 0.7948$, y cuando x_4 se utiliza con x_1 y x_3, se obtiene $R(\beta_4 \mid \beta_1, \beta_3) = 0.1855$. El valor de s^2 es 0.5979 para la combinación (x_1, x_2, x_3), y 0.7198 para la combinación (x_1, x_2, x_4). Como ningún valor f es significativo al nivel $\alpha = 0.05$, el modelo final de regresión sólo incluye las variables x_1 y x_3. Se encuentra que la ecuación de estimación es

$$\hat{y} = 20.1084 + 0.4136x_1 + 2.0253x_3,$$

y el coeficiente de determinación para este modelo es $R^2 = 0.9882$.

Aunque (x_1, x_3) es la combinación elegida mediante la regresión por etapas, no es necesariamente la combinación de dos variables que proporciona el valor más grande de R^2. De hecho, ya observamos que la combinación (x_2, x_3) da un valor de $R^2 = 0.9905$. Desde luego, el procedimiento por etapas nunca tomó en cuenta dicha combinación. Se podría plantear un argumento racional de que en realidad hay una diferencia despreciable en el desempeño entre esas dos ecuaciones de estimación, al menos en términos del

porcentaje de variación explicado. Sin embargo, es interesante observar que el procedimiento de eliminación hacia atrás proporciona la combinación (x_2, x_3) en la ecuación final (véase el ejercicio 12.49 en la página 494). ■

Resumen

La función principal de cada uno de los procedimientos explicados en esta sección consiste en exponer las variables a una metodología sistemática, diseñada para garantizar la inclusión final de las mejores combinaciones de las mismas. Es evidente que no es seguro que esto pase en todos los problemas y, por supuesto, es posible que la multicolinealidad sea tan extensa que no haya más alternativa que apoyarse en procedimientos de estimación diferentes de los mínimos cuadrados. Tales procedimientos de estimación se estudian en Myers (1990), listado en la bibliografía.

Los procedimientos secuenciales que se estudian aquí son tres de los muchos métodos de ese tipo que aparecen en la literatura y que están incluidos en diversos paquetes de regresión por computadora. Estos métodos fueron diseñados para ser eficientes en cuanto al cálculo pero, por supuesto, no proporcionan resultados para todos los subconjuntos posibles de variables. Debido a esto los procedimientos son más eficaces para conjuntos de datos que incluyen un **número grande de variables**. En el caso de los problemas de regresión que implican un número relativamente pequeño de variables, los paquetes modernos de cómputo para la regresión permiten el cálculo y resumen la información cuantitativa de todos los modelos para cada subconjunto posible de variables. En la sección 12.11 se proporcionan ilustraciones.

Elección de valores *P*

Como es de esperarse, la elección del modelo final con estos procedimientos podría depender en gran medida del valor *P* que se seleccione. Además, un procedimiento es más exitoso cuando es forzado a probar una gran cantidad de variables posibles. Por esta razón, cualquier procedimiento hacia delante es más útil cuando se utiliza un valor *P* relativamente grande. A esto se debe que algunos programas de cómputo empleen un valor *P* predeterminado de 0.50.

12.10 Estudio de los residuales y violación de las suposiciones (verificación del modelo)

Anteriormente en este capítulo se sugirió que los residuales, o errores en el ajuste de regresión, con frecuencia proporcionan información que puede ser muy valiosa para el analista de datos. Los $e_i = y_i = \hat{y}_i$, $i = 1, 2,..., n$, que son el equivalente numérico de los ϵ_i, los errores del modelo, a menudo revelan la posible violación de las suposiciones o la presencia de datos de puntos "sospechosos". Suponga que el vector $\mathbf{x}_i$ denota los valores de las variables regresoras que corresponden al *i*-ésimo punto de los datos, complementado por un 1 en la posición inicial. Es decir,

$$\mathbf{x}_i' = [1, x_{1i}, x_{2i}, \ldots, x_{ki}].$$

Considere la cantidad

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \qquad i = 1, 2, ..., n.$$

El lector debería notar que en la sección 12.5 se utilizó h_{ii} para calcular los intervalos de confianza de la respuesta media. Además de σ^2, h_{ii} representa la varianza del valor ajustado $\hat{y}_i$. Los valores h_{ii} son los elementos de la diagonal de la **matriz "SOMBRERO"**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

que desempeña un papel importante en cualquier estudio de residuales y en otros aspectos modernos del análisis de regresión (véase Myers, 1990, citado en la bibliografía). El término *matriz SOMBRERO* se deriva del hecho de que **H** genera las "*y* sombrero", o los valores ajustados cuando se multiplica por el vector **y** de respuestas observadas. Es decir, $\hat{\mathbf{y}} = \mathbf{Xb}$, por lo tanto,

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy},$$

donde $\hat{\mathbf{y}}$ es el vector cuyo i-ésimo elemento es $\hat{y}_i$.

Si se hacen las suposiciones acostumbradas de que los ϵ_i son independientes y están distribuidos normalmente, con media cero y varianza σ^2, las propiedades estadísticas de los residuales se establecen con facilidad. Entonces,

$$E(e_i) = E(y_i - \hat{y}_i) = 0 \quad \text{y} \quad \sigma^2_{e_i} = (1 - h_{ii})\sigma^2,$$

para $i = 1, 2,..., n$. (Para mayores detalles véase Myers, 1990). Es posible demostrar que los valores de la diagonal de la matriz SOMBRERO están acotados de acuerdo con la desigualdad

$$\frac{1}{n} \leq h_{ii} \leq 1.$$

Además, $\sum_{i=1}^{n} h_{ii} = k + 1$, el número de parámetros de la regresión. Como resultado, cualquier punto de los datos cuyo elemento diagonal SOMBRERO sea grande, es decir, esté muy por encima del valor promedio de $(k + 1)/n$, está en una posición dentro del conjunto de datos donde la varianza de $\hat{y}_i$ es relativamente grande y la varianza de un residuo es relativamente pequeña. Como resultado, el analista de datos puede tener una idea de qué tan grande puede ser un residuo antes de que su desviación de cero se pueda atribuir a algo distinto del azar. Muchos de los paquetes comerciales para computadora que permiten calcular la regresión producen el conjunto de **residuales estudentizados**.

Residuo estudentizado

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n$$

Aquí, cada residuo se **dividió entre una estimación de su desviación estándar** creando un estadístico *tipo t* diseñado para dar al analista una cantidad sin escala que proporcione información sobre el *tamaño* del residual. Además, a menudo los paquetes de cómputo comunes proporcionan valores de otro conjunto de residuales tipo estudentizados denominados **valores *R* de Student**.

Residual *R* de Student

$$t_i = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n,$$

donde s_{-i} es un estimador de la desviación estándar del error calculado con el ***i*-ésimo punto de los datos eliminado**.

Hay tres tipos de transgresiones de las suposiciones fáciles de detectar mediante el uso de los residuales o *gráficas de residuales*. Aunque las gráficas de los residuales brutos, los e_i, ayudan a esto, con frecuencia es más informativo graficar los residuales estudentizados. Las tres transgresiones son las siguientes:

1. Presencia de valores extremos
2. Varianza heterogénea del error
3. Especificación inadecuada del modelo

En el caso 1 elegimos definir un **valor extremo** como un punto de los datos que se desvía de la suposición común de que $E(\epsilon_i) = 0$ para un valor específico de i. Si hay una razón para creer que un punto de un dato específico es un valor extremo que ejerce una gran influencia sobre el modelo ajustado, r_i o t_i, esto podría estar informando algo. Es de esperarse que los valores R de Student sean más sensibles a los valores extremos que los valores r_i.

En realidad, en el caso de que $E(\varepsilon_i) = 0$, t_i es un valor de una variable aleatoria que sigue una distribución t con $n - 1 - (k + 1) = n - k - 2$ grados de libertad. Por consiguiente, es posible utilizar una prueba t de dos colas para proporcionar información con el fin de detectar si el punto i-ésimo es o no un valor extremo.

Aunque el estadístico R de Student t_i produce una prueba t exacta para detectar un valor extremo en una ubicación específica, la distribución t no se aplicaría para probar simultáneamente varios valores extremos en todas las ubicaciones. Como resultado, los residuales estudentizados o valores R de Student se deberían usar estrictamente como herramientas de diagnóstico *sin* un mecanismo de prueba de hipótesis formal. La implicación es que dichos estadísticos resaltan puntos de los datos en los que el error del ajuste es mayor de lo esperado por el azar. Los valores R de Student de gran magnitud sugieren la necesidad de "verificar" los datos con todos los recursos disponibles. La práctica de eliminar observaciones de conjuntos de datos de la regresión no debería llevarse a cabo de forma indiscriminada. (Para más información sobre el uso de los diagnósticos sobre valores extremos véase Myers, 1990, en la bibliografía).

Ilustración de la detección de valores extremos

Estudio de caso 12.1: **Método para capturar saltamontes.** En un experimento biológico, que fue efectuado en el Departamento de Entomología de Virginia Tech, se hicieron n ensayos experimentales con dos métodos diferentes para capturar saltamontes. Los métodos consistieron en la captura por caída de la red y la captura por barrido de la red. El número promedio de saltamontes atrapados con cada método se registró en un conjunto de cuadrantes del campo en una fecha determinada. También se registró una variable regresora adicional, la altura promedio de las plantas en los cuadrantes. Los datos experimentales aparecen en la tabla 12.10.

El objetivo consiste en estimar cuántos saltamontes se capturan empleando sólo el método del barrido de la red, que es menos costoso. Hay cierta preocupación por la validez del cuarto punto de los datos. La captura observada utilizando el método de caída de la red que se reportó parece inusualmente alta, dadas las demás condiciones, de hecho se pensó que la cifra podía ser errónea. Ajuste un modelo del tipo

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

para los 17 puntos de los datos y estudie los residuales para determinar si el punto 4 es un valor extremo.

Tabla 12.10: Conjunto de datos para el estudio de caso 12.1

Observación	Captura por caída de la red, y	Captura por barrido de la red, x_1	Altura de las plantas, x_2 (cm)
1	18.0000	4.15476	52.705
2	8.8750	2.02381	42.069
3	2.0000	0.15909	34.766
4	20.0000	2.32812	27.622
5	2.3750	0.25521	45.879
6	2.7500	0.57292	97.472
7	3.3333	0.70139	102.062
8	1.0000	0.13542	97.790
9	1.3333	0.12121	88.265
10	1.7500	0.10937	58.737
11	4.1250	0.56250	42.386
12	12.8750	2.45312	31.274
13	5.3750	0.45312	31.750
14	28.0000	6.68750	35.401
15	4.7500	0.86979	64.516
16	1.7500	0.14583	25.241
17	0.1333	0.01562	36.354

***Solución*:** Un paquete de cómputo generó el modelo de regresión ajustado

$$\hat{y} = 3.6870 + 4.1050x_1 - 0.0367x_2$$

junto con los estadísticos $R^2 = 0.9244$ y $s^2 = 5.580$. También se obtuvieron los residuales y otra información de diagnóstico que fueron registrados en la tabla 12.11.

Como se esperaba, el residual en la cuarta ubicación parece inusualmente grande, a saber, 7.769. La cuestión fundamental aquí es si este residual es más grande de lo que se esperaría debido al azar. El error estándar del residual para el punto 4 es 2.209. El valor R de Student t_4 que se obtuvo es 9.9315. Al considerarlo como el valor de una variable aleatoria que tiene una distribución t con 13 grados de libertad, se concluiría sin duda que el residuo de la cuarta observación se estima algo mayor que 0, y que la medición del presunto error es apoyada por el estudio de los residuales. Observe que ningún otro valor de los residuales proporciona un valor R de Student que sea motivo de alarma. ■

Gráfica de los residuales para el estudio de caso 12.1

En el capítulo 11 estudiamos con cierto detalle la utilidad de graficar los residuos en el análisis de regresión. Es frecuente que con base en dichas gráficas se detecte la violación de las suposiciones del modelo. En la regresión múltiple en ocasiones es útil graficar la probabilidad normal de los residuales o los residuales en comparación con $\hat{y}$. Sin embargo, a menudo es preferible graficar los residuales estudentizados.

Recuerde que la preferencia por los residuales estudentizados sobre los residuales ordinarios para propósitos de graficación se debe a que, como la varianza de i-ésimo

Tabla 12.11: Información sobre los residuales para el conjunto de datos del estudio de caso 12.1

Obs.	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	h_{ii}	$s\sqrt{1-h_{ii}}$	r_i	t_i
1	18.000	18.809	−0.809	0.2291	2.074	−0.390	−0.3780
2	8.875	10.452	−1.577	0.0766	2.270	−0.695	−0.6812
3	2.000	3.065	−1.065	0.1364	2.195	−0.485	−0.4715
4	20.000	12.231	7.769	0.1256	2.209	3.517	9.9315
5	2.375	3.052	−0.677	0.0931	2.250	−0.301	−0.2909
6	2.750	2.464	0.286	0.2276	2.076	0.138	0.1329
7	3.333	2.823	0.510	0.2669	2.023	0.252	0.2437
8	1.000	0.656	0.344	0.2318	2.071	0.166	0.1601
9	1.333	0.947	0.386	0.1691	2.153	0.179	0.1729
10	1.750	1.982	−0.232	0.0852	2.260	−0.103	−0.0989
11	4.125	4.442	−0.317	0.0884	2.255	−0.140	−0.1353
12	12.875	12.610	0.265	0.1152	2.222	0.119	0.1149
13	5.375	4.383	0.992	0.1339	2.199	0.451	0.4382
14	28.000	29.841	−1.841	0.6233	1.450	−1.270	−1.3005
15	4.750	4.891	−0.141	0.0699	2.278	−0.062	−0.0598
16	1.750	3.360	−1.610	0.1891	2.127	−0.757	−0.7447
17	0.133	2.418	−2.285	0.1386	2.193	−1.042	−1.0454

residuo depende del i-ésimo elemento en la diagonal SOMBRERO, las varianzas de los residuos diferirán si hay dispersión en las diagonales SOMBRERO. Así, es probable que la apariencia de una gráfica de residuales sugiera heterogeneidad debido a que los propios residuales no se comportan, en general, de manera ideal. El propósito de utilizar residuales estudentizados es proporcionar un tipo de *estandarización*. Es evidente que si se conociera σ, en condiciones ideales, es decir, en las que el modelo fuera correcto y la varianza homogénea, se tendría

$$E\left(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}\right) = 0 \quad \text{y} \quad \text{Var}\left(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}\right) = 1.$$

De manera que los residuales estudentizados producen un conjunto de estadísticos que en condiciones ideales se comportan en forma estándar. La figura 12.5 presenta una gráfica con los valores ***R* de Student** para los datos de los saltamontes del estudio de caso 12.1. Advierta que el valor para la observación 4 se destaca de los demás. La gráfica *R* de Student se generó con el programa *SAS*. La gráfica presenta los residuales en comparación con los valores $\hat{y}$.

Verificación de la normalidad

El lector debe recordar, de acuerdo con lo que se estudió en el capítulo 11, la importancia de verificar la normalidad utilizando una gráfica de probabilidad normal. La misma recomendación es válida para el caso de la regresión lineal múltiple. Las gráficas de probabilidad normal se pueden generar utilizando software estándar para regresión. Sin embargo, como ya se indicó, éstas pueden ser más eficaces si se usan residuales estudentizados o valores *R* de Student en vez de residuales comunes.

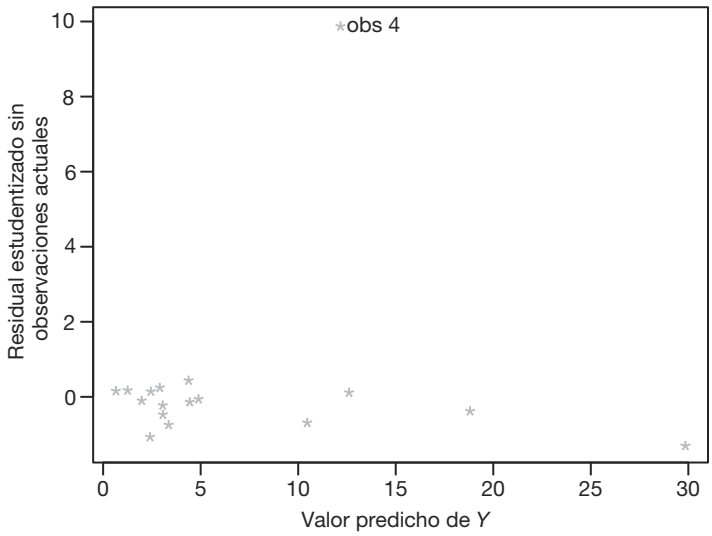


Figura 12.5: Valores *R* de Student graficados en comparación con los valores predichos para los datos de los saltamontes del estudio de caso 12.1.

12.11 Validación cruzada, C_p, y otros criterios para la selección del modelo

Para muchos problemas de regresión el experimentador debe elegir entre varios modelos alternativos o formas de modelo que se desarrollan a partir del mismo conjunto de datos. Con mucha frecuencia se requiere el modelo que predice o estima mejor la respuesta media. El experimentador debe tomar en cuenta los tamaños relativos de los valores de s^2 para los posibles modelos y, sin duda, la naturaleza general de los intervalos de confianza sobre la respuesta media. También se debe considerar lo bien que el modelo predice los valores de la respuesta que **no se hayan utilizado para construir los posibles modelos**. Los modelos deben estar sujetos a **validación cruzada**. Entonces, lo que se requiere son los errores de la validación cruzada en lugar de los errores del ajuste. Estos errores en la predicción son los **residuales PRESS**.

$$\delta_i = y_i - \hat{y}_{i,-i}, \quad \text{i} = 1, 2, \ldots, n,$$

donde $\hat{y}_{i,-i}$ es la predicción del *i*-ésimo punto de los datos por medio de un modelo que no utiliza el *i*-ésimo punto en el cálculo de los coeficientes. Estos residuales PRESS se calculan mediante la fórmula

$$\delta_i = \frac{e_i}{1 - h_{ii}}, \quad i = 1\ ,2, \ldots, n.$$

(La derivación se encuentra en Myers, 1990).

Uso del estadístico PRESS

La motivación para utilizar PRESS y la utilidad de los residuales PRESS es muy fácil de entender. El propósito de extraer o *separar* puntos de datos, uno a la vez, consiste en

permitir el empleo de metodologías separadas para ajustar y evaluar un modelo específico. Para evaluar un modelo la "$-i$" indica que el residual PRESS proporciona un error de predicción donde la observación que se predice es *independiente del ajuste del modelo*.

Los criterios que utilizan los residuales PRESS son dados por

$$\sum_{i=1}^{n} |\delta_i| \quad \text{y} \quad \text{PRESS} = \sum_{i=1}^{n} \delta_i^2.$$

(El término **PRESS** es un acrónimo que se forma con las iniciales de los términos de la frase en inglés *prediction sum of squares,* que se traduce como **suma de cuadrados de predicción**). Se sugiere que se utilicen ambos criterios. Es posible que PRESS sea dominado por uno o algunos residuales PRESS grandes. Es evidente que el criterio sobre $\sum_{i=1}^{n}|\delta_i|$ es menos sensible a un número pequeño de valores grandes.

Además del estadístico PRESS en sí, el analista puede simplemente calcular un estadístico similar a R^2 que refleje el desempeño de la predicción. Con frecuencia a este estadístico se le denomina R^2_{pred} y se calcula como sigue:

R^2 de predicción Dado un modelo ajustado con valor específico para PRESS, R^2_{pred} es dado por

$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Observe que R^2_{pred} es tan sólo el estadístico común R^2 donde la *SCE* fue reemplazada por el estadístico PRESS.

En el siguiente estudio de caso se proporciona un ejemplo en el que se ajustan muchos posibles modelos a un conjunto de datos y se elige el mejor de ellos. No se emplean los procedimientos secuenciales descritos en la sección 12.9. En vez de eso se ilustra el papel que desempeñan los residuales PRESS y otros valores estadísticos cuando se trata de seleccionar la mejor ecuación de regresión.

Estudio de caso 12.2: **Patada de fútbol.** La fuerza de las piernas es un requisito necesario para que un pateador tenga éxito en el fútbol americano. Una medida de la calidad de una buena patada es el "tiempo de vuelo" del balón, es decir, el tiempo que el balón se mantiene en el aire antes de ser atrapado por el regresador de patadas. Para determinar cuáles factores de la fuerza de las piernas influyen en el tiempo de vuelo y desarrollar un modelo empírico para predecir esta respuesta, el Departamento de Salud, Educación Física y Recreación de Virginia Tech llevó a cabo un estudio sobre *La relación entre variables seleccionadas de desempeño físico y la capacidad de despejes en el fútbol*. Se eligieron 13 pateadores para el experimento y cada uno pateó 10 veces el balón. En la tabla 12.12 aparece el registro del tiempo de vuelo promedio, junto con las medidas de fuerza usadas en el análisis.

Cada variable regresora se define como sigue:

1. **FPD**, fuerza de la pierna derecha (libras)
2. **FPI**, fuerza de la pierna izquierda (libras)
3. **FTD**, flexibilidad muscular del tendón derecho (grados)
4. **FTI**, flexibilidad muscular del tendón izquierdo (grados)

5. Potencia, fuerza general de las piernas (pie-libras)

Determine el modelo más adecuado para predecir el tiempo de vuelo.

Tabla 12.12: Datos para el estudio de caso 12.2.

Pateador	Tiempo de vuelo, y (seg)	FPD, x_1	FPI, x_2	FTD, x_3	FTI, x_4	Potencia, x_5
1	4.75	170	170	106	106	240.57
2	4.07	140	130	92	93	195.49
3	4.04	180	170	93	78	152.99
4	4.18	160	160	103	93	197.09
5	4.35	170	150	104	93	266.56
6	4.16	150	150	101	87	260.56
7	4.43	170	180	108	106	219.25
8	3.20	110	110	86	92	132.68
9	3.02	120	110	90	86	130.24
10	3.64	130	120	85	80	205.88
11	3.68	120	140	89	83	153.92
12	3.60	140	130	92	94	154.64
13	3.85	160	150	95	95	240.57

***Solución*:** Al buscar el mejor modelo posible para predecir el tiempo de vuelo se obtuvo la información de la tabla 12.13 utilizando un paquete de cómputo para regresión. Los modelos están clasificados en orden ascendente con respecto a los valores del estadístico PRESS. Esta presentación brinda información suficiente acerca de todos los modelos posibles con el fin de permitir que el usuario elimine algunos de ellos. Al parecer, el mejor modelo para predecir el tiempo de vuelo para los pateadores es el que contiene a x_2 y x_5 (FPI y potencia), denotadas por x_2x_5. Asimismo, observe que todos los modelos con valores bajos de PRESS, de s^2, de $\sum_{i=1}^{n}|\delta_i|$, y con valores altos de R^2, contienen esas dos variables. Para obtener información de los residuales de la regresión ajustada

$$\hat{y}_i = b_0 + b_2x_{2i} + b_5x_{5i},$$

se generaron los residuales y los residuales PRESS. El modelo de predicción real (véase el ejercicio 12.47 de la página 494) es dado por

$$\hat{y} = 1.10765 + 0.01370x_2 + 0.00429x_5.$$

En la tabla 12.14 se listan los residuales, los valores de la diagonal testada y los valores PRESS.

Observe el ajuste relativamente bueno de los modelos de regresión con dos variables para los datos. Los residuales PRESS reflejan la capacidad de la ecuación de regresión para predecir el tiempo de vuelo si se hicieran predicciones independientes. Por ejemplo, para el pateador número 4 el tiempo de vuelo de 4.180 tendría un error de predicción de 0.039 si se construyera el modelo usando a los 12 pateadores restantes. Para este modelo el error promedio de la predicción, o error de validación cruzada, es

$$\frac{1}{13}\sum_{i=1}^{n}|\delta_i| = 0.1489 \text{ segundos},$$

Tabla 12.13: Comparación de diferentes modelos de regresión

Modelo	s^2	$\sum \lvert\delta_i\rvert$	**PRESS**	R^2
x_2x_5	0.036907	1.93583	0.54683	0.871300
$x_1x_2x_5$	0.041001	2.06489	0.58998	0.871321
$x_2x_4x_5$	0.037708	2.18797	0.59915	0.881658
$x_2x_3x_5$	0.039636	2.09553	0.66182	0.875606
$x_1x_2x_4x_5$	0.042265	2.42194	0.67840	0.882093
$x_1x_2x_3x_5$	0.044578	2.26283	0.70958	0.875642
$x_2x_3x_4x_5$	0.042421	2.55789	0.86236	0.881658
$x_1x_3x_5$	0.053664	2.65276	0.87325	0.831580
$x_1x_4x_5$	0.056279	2.75390	0.89551	0.823375
x_1x_5	0.059621	2.99434	0.97483	0.792094
x_2x_3	0.056153	2.95310	0.98815	0.804187
x_1x_3	0.059400	3.01436	0.99697	0.792864
$x_1x_2x_3x_4x_5$	0.048302	2.87302	1.00920	0.882096
x_2	0.066894	3.22319	1.04564	0.743404
x_3x_5	0.065678	3.09474	1.05708	0.770971
x_1x_2	0.068402	3.09047	1.09726	0.761474
x_3	0.074518	3.06754	1.13555	0.714161
$x_1x_3x_4$	0.065414	3.36304	1.15043	0.794705
$x_2x_3x_4$	0.062082	3.32392	1.17491	0.805163
x_2x_4	0.063744	3.59101	1.18531	0.777716
$x_1x_2x_3$	0.059670	3.41287	1.26558	0.812730
x_3x_4	0.080605	3.28004	1.28314	0.718921
x_1x_4	0.069965	3.64415	1.30194	0.756023
x_1	0.080208	3.31562	1.30275	0.692334
$x_1x_3x_4x_5$	0.059169	3.37362	1.36867	0.834936
$x_1x_2x_4$	0.064143	3.89402	1.39834	0.798692
$x_3x_4x_5$	0.072505	3.49695	1.42036	0.772450
$x_1x_2x_3x_4$	0.066088	3.95854	1.52344	0.815633
x_5	0.111779	4.17839	1.72511	0.571234
x_4x_5	0.105648	4.12729	1.87734	0.631593
x_4	0.186708	4.88870	2.82207	0.283819

que es pequeño comparado con el tiempo de vuelo promedio para los 13 pateadores. ∎

En la sección 12.9 indicamos que a menudo es aconsejable utilizar todos los subconjuntos posibles de regresión cuando se busca el mejor modelo. La mayoría de los programas comerciales de cómputo para estadística contienen una rutina de *todas las regresiones posibles*. Tales algoritmos calculan diversos criterios para todos los subconjuntos de términos del modelo. Es evidente que criterios como R^2, s^2 y PRESS son razonables para elegir entre subconjuntos de candidatos. Otro estadístico muy popular y útil, en particular para las ciencias físicas e ingeniería, es el estadístico C_p, que se describe a continuación.

Tabla 12.14: Residuales PRESS

Pateador	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	h_{ii}	δ_i
1	4.750	4.470	0.280	0.198	0.349
2	4.070	3.728	0.342	0.118	0.388
3	4.040	4.094	−0.054	0.444	−0.097
4	4.180	4.146	0.034	0.132	0.039
5	4.350	4.307	0.043	0.286	0.060
6	4.160	4.281	−0.121	0.250	−0.161
7	4.430	4.515	−0.085	0.298	−0.121
8	3.200	3.184	0.016	0.294	0.023
9	3.020	3.174	−0.154	0.301	−0.220
10	3.640	3.636	0.004	0.231	0.005
11	3.680	3.687	−0.007	0.152	−0.008
12	3.600	3.553	0.047	0.142	0.055
13	3.850	4.196	−0.346	0.154	−0.409

El estadístico C_p

Muy a menudo la selección del modelo más adecuado implica tomar en cuenta muchas cosas. Evidentemente el número de términos del modelo es importante; el tema de la parsimonia no debe ignorarse. Por otro lado, el analista no debe sentirse satisfecho con un modelo demasiado simple hasta el punto de una simplificación excesiva. En este sentido, un estadístico único que implica un compromiso aceptable es C_p. (Véase Mallows, 1973, en la bibliografía).

El estadístico C_p apela de forma muy adecuada al sentido común y se desarrolla tomando en cuenta el equilibrio apropiado entre el sesgo excesivo en que se incurre cuando se subajusta, es decir, cuando se eligen muy pocos términos para el modelo; y la varianza excesiva de la predicción que se genera cuando se sobreajusta, o sea cuando hay redundancias en el modelo. El estadístico C_p es una función simple del número total de parámetros en el posible modelo y la media cuadrada del error s^2.

Aquí no presentaremos el desarrollo completo del estadístico C_p. (Para mayores detalles se recomienda consultar a Myers, 1990, listado en la bibliografía). El C_p para un subconjunto particular de modelos es *un estimado* de lo siguiente:

$$\Gamma_{(p)} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \text{Var}(\hat{y}_i) + \frac{1}{\sigma^2}\sum_{i=1}^{n} (\text{Sesgo}\,\hat{y}_i)^2.$$

Se descubre que, bajo las suposiciones estándar de los mínimos cuadrados que se indicaron con anterioridad en este capítulo, y asumiendo que el modelo "verdadero" es aquel que contiene todas las posibles variables,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \text{Var}(\hat{y}_i) = p \qquad \text{(número de parámetros en el posible modelo)}$$

(véase el ejercicio de repaso 12.63) y un estimado no sesgado de

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\text{Sesgo}\,\hat{y}_i)^2 \text{ es dado por } \frac{1}{\sigma^2}\sum_{i=1}^{n}(\widehat{\text{Sesgo}}\,\hat{y}_i)^2 = \frac{(s^2-\sigma^2)(n-p)}{\sigma^2}.$$

En las ecuaciones anteriores s^2 es el cuadrado medio del error para el posible modelo y σ^2 es la varianza del error de la población. Así, si asumimos que se dispone de algún estimado $\hat{\sigma}^2$ para σ^2, entonces C_p es dado por la siguiente ecuación:

Estadístico C_p

$$C_p = p + \frac{(s^2-\hat{\sigma}^2)(n-p)}{\hat{\sigma}^2},$$

donde p es el número de parámetros en el modelo, s^2 es el cuadrado medio del error para el modelo candidato y $\hat{\sigma}^2$ es un estimador de σ^2.

Es evidente que el científico debería adoptar modelos con valores pequeños de C_p. El lector observará que, a diferencia del estadístico PRESS, C_p carece de una escala. Además, se puede obtener cierta información acerca de qué tan adecuado es un posible modelo observando su valor de C_p. Por ejemplo, $C_p > p$ indica que un modelo está sesgado debido a que está subajustado, mientras que $C_p \approx p$ indica un modelo razonable.

Con frecuencia hay confusión respecto a la procedencia de $\hat{\sigma}^2$ en la fórmula para C_p. Es evidente que el científico o ingeniero no tienen acceso a la cantidad σ^2 de la población. En aplicaciones donde se dispone de corridas repetidas, digamos en situaciones de diseño experimental, se dispone de un estimado de σ^2 independiente del modelo (véase los capítulos 11 y 15). Sin embargo, la mayoría de paquetes de cómputo utilizan $\hat{\sigma}^2$ como el *cuadrado medio del error del modelo más completo*. Evidentemente, si éste no es un buen estimado, la parte de sesgo del estadístico C_p puede ser negativa. Por consiguiente, C_p puede ser menor que p.

Ejemplo 12.12: Considere el conjunto de datos de la tabla 12.15, los cuales reflejan el interés de un fabricante de grava asfáltica en la relación que existe entre las ventas durante un año específico y los factores que influyen en ellas. (Los datos fueron tomados de Kutner *et al.,* 2004, véase la bibliografía).

En los subconjuntos de modelos posibles, hay tres que revisten interés especial. Estos tres son los de x_2x_3, $x_1x_2x_3$ y $x_1x_2x_3x_4$. A continuación se presenta la información pertinente para comparar los tres modelos. Para ayudar a la toma de decisiones incluimos los estadísticos PRESS de los tres modelos.

Modelo	R^2	R^2_{pred}	s^2	PRESS	C_p
x_2x_3	0.9940	0.9913	44.5552	782.1896	11.4013
$x_1x_2x_3$	0.9970	0.9928	24.7956	643.3578	3.4075
$x_1x_2x_3x_4$	0.9971	0.9917	26.2073	741.7557	5.0

A partir de la información de la tabla parece claro que el modelo $x_1x_2x_3$ es mejor que los otros dos. Observe que para el modelo completo $C_p = 5.0$. Esto ocurre porque la *parte de sesgo* es igual a cero y $\hat{\sigma}^2 = 26.2073$ es el cuadrado medio del error del modelo completo. ∎

La figura 12.6 es una salida de resultados de la función PROC REG del *SAS,* la cual muestra información sobre todas las regresiones posibles. A partir de ella es posible hacer comparaciones de otros modelos con (x_1, x_2, x_3). Observe que (x_1, x_2, x_3) parece muy bueno en comparación con todos los demás modelos.

Como verificación final del modelo (x_1, x_2, x_3), la figura 12.7 presenta una gráfica de probabilidad normal de los residuales del modelo.

Tabla 12.15: Datos para el ejemplo 12.12

Distrito	Cuentas promocionales, x_1	Cuentas activas, x_2	Marcas en competencia, x_3	Potencial, x_4	Ventas, y (miles)
1	5.5	31	10	8	$ 79.3
2	2.5	55	8	6	200.1
3	8.0	67	12	9	163.2
4	3.0	50	7	16	200.1
5	3.0	38	8	15	146.0
6	2.9	71	12	17	177.7
7	8.0	30	12	8	30.9
8	9.0	56	5	10	291.9
9	4.0	42	8	4	160.0
10	6.5	73	5	16	339.4
11	5.5	60	11	7	159.6
12	5.0	44	12	12	86.3
13	6.0	50	6	6	237.5
14	5.0	39	10	4	107.2
15	3.5	55	10	4	155.0

Dependent Variable: sales

Number in Model	C(p)	R-Square	Adjusted R-Square	MSE	Variables in Model
3	3.4075	0.9970	0.9961	24.79560	x1 x2 x3
4	5.0000	0.9971	0.9959	26.20728	x1 x2 x3 x4
2	11.4013	0.9940	0.9930	44.55518	x2 x3
3	13.3770	0.9940	0.9924	48.54787	x2 x3 x4
3	1053.643	0.6896	0.6049	2526.96144	x1 x3 x4
2	1082.670	0.6805	0.6273	2384.14286	x3 x4
2	1215.316	0.6417	0.5820	2673.83349	x1 x3
1	1228.460	0.6373	0.6094	2498.68333	x3
3	1653.770	0.5140	0.3814	3956.75275	x1 x2 x4
2	1668.699	0.5090	0.4272	3663.99357	x1 x2
2	1685.024	0.5042	0.4216	3699.64814	x2 x4
1	1693.971	0.5010	0.4626	3437.12846	x2
2	3014.641	0.1151	-.0324	6603.45109	x1 x4
1	3088.650	0.0928	0.0231	6248.72283	x4
1	3364.884	0.0120	-.0640	6805.59568	x1

Figura 12.6: Salida de resultados del *SAS* de todos los subconjuntos posibles sobre los datos de las ventas para el ejemplo 12.12.

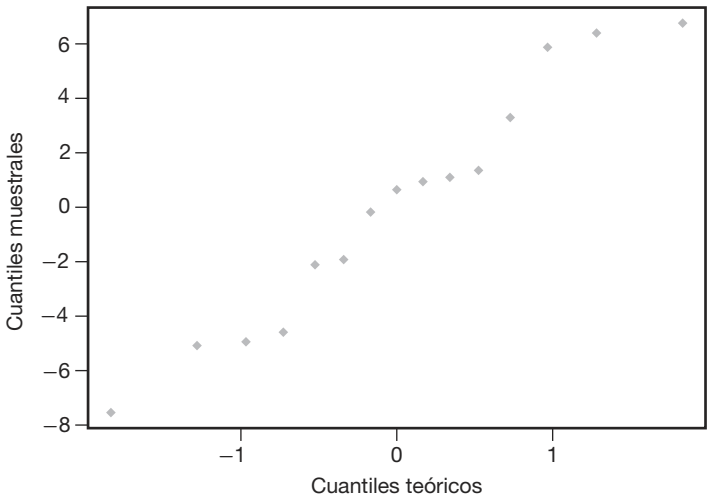


Figura 12.7: Gráfica de la probabilidad normal de los residuales, utilizando el modelo $x_1x_2x_3$ para el ejemplo 12.12.

Ejercicios

12.47 Considere los datos sobre el "tiempo de vuelo" de los pateadores que se presentaron en el estudio de caso 12.2 y, utilizando sólo las variables x_2 y x_3.

a) Verifique la ecuación de regresión que se presenta en la página 489.
b) Prediga el tiempo de vuelo para un pateador con FPI = 180 libras y potencia = 260 pie-libras.
c) Construya un intervalo de confianza de 95% para el tiempo de vuelo promedio de un pateador con FPI = 180 libras y potencia = 260 pies-libras.

12.48 Para los datos del ejercicio 12.15 de la página 452 utilice las técnicas de

a) *selección hacia delante* a un nivel de significancia de 0.05 para elegir un modelo de regresión lineal;
b) *eliminación hacia atrás* a un nivel de significancia de 0.05 para seleccionar un modelo de regresión lineal;
c) *regresión por etapas* a un nivel de significancia de 0.05 para escoger un modelo de regresión lineal.

12.49 Emplee las técnicas de *eliminación hacia atrás* con $\alpha = 0.05$ para elegir una ecuación de predicción para los datos de la tabla 12.8.

12.50 Para los datos de los pateadores del estudio de caso 12.2 también se registró una respuesta adicional, la "distancia de la patada". Los siguientes son los valores de distancia promedio para cada uno de los 13 pateadores:

a) Utilice los datos de distancia en lugar de los de tiempo de vuelo para estimar un modelo de regresión lineal múltiple del tipo

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

que permita predecir la distancia de la patada.
b) Utilice la regresión por etapas a un nivel de significancia de 0.10 para seleccionar una combinación de variables.
c) Genere valores para s^2, R^2, PRESS y $\sum_{i=1}^{n}|\delta_i|$ para el conjunto completo de 31 modelos. Utilice esta información para determinar la mejor combinación de variables para predecir la distancia de la patada.
d) Para el modelo final que seleccione, grafique los residuales estandarizados en comparación con Y y elabore una gráfica de probabilidad normal de los residuales ordinarios. Haga comentarios al respecto.

Pateador	Distancia, y (pies)
1	162.50
2	144.00
3	147.50
4	163.50
5	192.00
6	171.75
7	162.00
8	104.93
9	105.67
10	117.59
11	140.25
12	150.17
13	165.16

12.51 El siguiente es un conjunto de datos para y, la cantidad de dinero (en miles de dólares) aportado a la asociación de exalumnos del Virginia Tech por la generación de 1960; y para x, el número de años que han transcurrido desde la graduación:

y	x	y	x
812.52	1	2755.00	11
822.50	2	4390.50	12
1211.50	3	5581.50	13
1348.00	4	5548.00	14
1301.00	8	6086.00	15
2567.50	9	5764.00	16
2526.50	10	8903.00	17

a) Ajuste un modelo de regresión del tipo

$$\mu_{Y|x} = \beta_0 + \beta_1 x.$$

b) Ajuste un modelo cuadrático del tipo

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_{11} x^2.$$

c) Determine cuál de los modelos de los incisos *a*) o *b*) es preferible. Utilice s^2, R^2 y los residuales PRESS para sustentar su decisión.

12.52 Para el modelo del ejercicio 12.50*a*) pruebe la hipótesis

$$H_0: \beta_4 = 0,$$
$$H_1: \beta_4 \neq 0.$$

Utilice un valor P para su conclusión.

12.53 Para el modelo cuadrático del ejercicio 12.51*b*) proporcione estimados de las varianzas y las covarianzas de los estimados de β_1 y β_{11}.

12.54 Un cliente del Departamento de Ingeniería Mecánica se acercó al Centro de Consulta de Virginia Tech para que lo ayudaran a analizar un experimento sobre motores con turbina de gas. Se midieron varias salidas del voltaje de los motores con distintas combinaciones de velocidad de las aspas y de la extensión de los sensores. Los datos son los siguientes:

y (voltios)	Velocidad, x_1 (pulg/seg)	Extensión, x_2 (pulg)
1.95	6336	0.000
2.50	7099	0.000
2.93	8026	0.000
1.69	6230	0.000
1.23	5369	0.000
3.13	8343	0.000
1.55	6522	0.006
1.94	7310	0.006
2.18	7974	0.006
2.70	8501	0.006
1.32	6646	0.012
1.60	7384	0.012
1.89	8000	0.012
2.15	8545	0.012
1.09	6755	0.018
1.26	7362	0.018
1.57	7934	0.018
1.92	8554	0.018

a) Ajuste una regresión lineal múltiple a los datos.

b) Calcule las pruebas t sobre los coeficientes. Proporcione valores P.

c) Haga comentarios sobre la calidad del modelo ajustado.

12.55 La blancura del rayón es un factor importante para los científicos que estudian la calidad de las telas. La blancura se ve afectada por la calidad de la pulpa y otras variables de procesamiento. Algunas de las variables son la temperatura del baño con ácido, °C (x_1); la concentración del ácido en cascada, % (x_2); la temperatura del agua, °C (x_3); la concentración del sulfuro, % (x_4); la cantidad del blanqueador de cloro, lb/min (x_5) y la temperatura de terminado de la tela, °C (x_6). A continuación se proporciona un conjunto de datos de especímenes de rayón. La respuesta, y, es la medida de la blancura.

y	x_1	x_2	x_3	x_4	x_5	x_6
88.7	43	0.211	85	0.243	0.606	48
89.3	42	0.604	89	0.237	0.600	55
75.5	47	0.450	87	0.198	0.527	61
92.1	46	0.641	90	0.194	0.500	65
83.4	52	0.370	93	0.198	0.485	54
44.8	50	0.526	85	0.221	0.533	60
50.9	43	0.486	83	0.203	0.510	57
78.0	49	0.504	93	0.279	0.489	49
86.8	51	0.609	90	0.220	0.462	64
47.3	51	0.702	86	0.198	0.478	63
53.7	48	0.397	92	0.231	0.411	61
92.0	46	0.488	88	0.211	0.387	88
87.9	43	0.525	85	0.199	0.437	63
90.3	45	0.486	84	0.189	0.499	58
94.2	53	0.527	87	0.245	0.530	65
89.5	47	0.601	95	0.208	0.500	67

a) Utilice los criterios *CME*, c_p y PRESS para obtener el "mejor" modelo de todos los subconjuntos de los modelos.

b) Grafique los residuales estandarizados en comparación con Y y dibuje una gráfica de probabilidad normal de los residuales para el "mejor" modelo. Comente sus resultados.

12.56 En un esfuerzo para modelar las remuneraciones de los ejecutivos en el año 1979 se seleccionaron 33 empresas y se recabaron datos acerca de las remuneraciones, las ventas, las utilidades y el empleo. Se reunieron los siguientes datos para el año 1979.

Empresa	Remuneraciones, y (miles)	Ventas, x_1, (millones)	Utilidades, x_2 (millones)	Empleo, x_3
1	$450	$4600.6	$128.1	48,000
2	387	9255.4	783.9	55,900
3	368	1526.2	136.0	13,783
4	277	1683.2	179.0	27,765
5	676	2752.8	231.5	34,000
6	454	2205.8	329.5	26,500
7	507	2384.6	381.8	30,800
8	496	2746.0	237.9	41,000
9	487	1434.0	222.3	25,900

(cont.)

Empresa	Remuneraciones, y (miles)	Ventas, x_1, (millones)	Utilidades, x_2 (millones)	Empleo, x_3
10	$383	$470.6	$63.7	8600
11	311	1508.0	149.5	21,075
12	271	464.4	30.0	6874
13	524	9329.3	577.3	39,000
14	498	2377.5	250.7	34,300
15	343	1174.3	82.6	19,405
16	354	409.3	61.5	3586
17	324	724.7	90.8	3905
18	225	578.9	63.3	4139
19	254	966.8	42.8	6255
20	208	591.0	48.5	10,605
21	518	4933.1	310.6	65,392
22	406	7613.2	491.6	89,400
23	332	3457.4	228.0	55,200
24	340	545.3	54.6	7800
25	698	22,862.8	3011.3	337,119
26	306	2361.0	203.0	52,000
27	613	2614.1	201.0	50,500
28	302	1013.2	121.3	18,625
29	540	4560.3	194.6	97,937
30	293	855.7	63.4	12,300
31	528	4211.6	352.1	71,800
32	456	5440.4	655.2	87,700
33	417	1229.9	97.5	14,600

Considere el modelo

$$y_i = \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} + \epsilon_i, \qquad i = 1, 2, \ldots, 33.$$

a) Ajuste la regresión con el modelo anterior.
b) ¿Un modelo con un subconjunto de variables es preferible al modelo completo?

12.57 La resistencia a la tracción de una unión de alambre es una característica importante. La siguiente tabla brinda información sobre la resistencia a la tracción, y, la altura del molde, x_1, la altura del perno, x_2, la altura del lazo, x_3, la longitud del alambre, x_4, el ancho de la unión sobre el molde, x_5 y el ancho del molde sobre el perno, x_6. (Datos tomados de Myers, Montgomery y Anderson-Cook, 2009).

a) Ajuste un modelo de regresión usando todas las variables independientes.
b) Utilice la regresión por etapas a un nivel de significancia de entrada de 0.25 y un nivel de significancia de eliminación de 0.05. Proporcione el modelo final.
c) Utilice todos los modelos de regresión posibles y calcule R^2, C_p, s^2 y R^2 ajustada para todos los modelos.
d) Proporcione el modelo final.
e) Para el modelo del inciso *d*) grafique los residuos estudentizados (o la R de Student) y haga comentarios al respecto.

y	x_1	x_2	x_3	x_4	x_5	x_6
8.0	5.2	19.6	29.6	94.9	2.1	2.3
8.3	5.2	19.8	32.4	89.7	2.1	1.8
8.5	5.8	19.6	31.0	96.2	2.0	2.0
8.8	6.4	19.4	32.4	95.6	2.2	2.1
9.0	5.8	18.6	28.6	86.5	2.0	1.8
9.3	5.2	18.8	30.6	84.5	2.1	2.1
9.3	5.6	20.4	32.4	88.8	2.2	1.9
9.5	6.0	19.0	32.6	85.7	2.1	1.9
9.8	5.2	20.8	32.2	93.6	2.3	2.1
10.0	5.8	19.9	31.8	86.0	2.1	1.8
10.3	6.4	18.0	32.6	87.1	2.0	1.6
10.5	6.0	20.6	33.4	93.1	2.1	2.1
10.8	6.2	20.2	31.8	83.4	2.2	2.1
11.0	6.2	20.2	32.4	94.5	2.1	1.9
11.3	6.2	19.2	31.4	83.4	1.9	1.8
11.5	5.6	17.0	33.2	85.2	2.1	2.1
11.8	6.0	19.8	35.4	84.1	2.0	1.8
12.3	5.8	18.8	34.0	86.9	2.1	1.8
12.5	5.6	18.6	34.2	83.0	1.9	2.0

12.58 Para el ejercicio 12.57 pruebe H_0: $\beta_1 = \beta_6 = 0$. Proporcione valores P y comente al respecto.

12.59 En el ejercicio 12.28 de la página 462 se tienen los siguientes datos sobre el desgaste de un cojinete:

y (desgaste)	x_1 (viscosidad del aceite)	x_2 (carga)
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

a) Puede considerar el siguiente modelo para describir los datos:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

para $i = 1, 2,..., 6$. El término x_1x_2 es una "interacción". Ajuste este modelo y estime los parámetros.
b) Utilice los modelos (x_1), (x_1, x_2), (x_2), (x_1, x_2, x_1x_2) y calcule PRESS, C_p, y s^2 para determinar el "mejor" modelo.

12.12 Modelos especiales no lineales para condiciones no ideales

En gran parte del material anterior de este capítulo y en el del capítulo 11 nos hemos beneficiado mucho de la suposición de que los errores del modelo, los ϵ_i, son normales,

con media igual a cero y varianza constante σ^2. Sin embargo, en la vida real hay muchas situaciones en las cuales es evidente que la respuesta no es normal. Por ejemplo, existe una gran cantidad de aplicaciones en las que la **respuesta es binaria** (0 o 1), por lo que su naturaleza es de Bernoulli. En las ciencias sociales un problema podría ser el de desarrollar un modelo que prediga si un individuo representa riesgos para un crédito (0 o 1), en función de ciertos regresores socioeconómicos, como sus ingresos, edad, género y nivel de escolaridad. En una prueba biomédica para un fármaco a menudo se observa si el paciente responde o no de manera favorable a éste, en tanto que los regresores podrían incluir la dosis y factores biológicos como la edad, el peso y la presión sanguínea. Nuevamente la respuesta es de naturaleza binaria. También abundan las aplicaciones en las áreas de manufactura en que ciertos factores controlables influyen en el hecho de que un artículo fabricado esté **o no defectuoso**.

Un segundo tipo de aplicación que no es normal y del que haremos una mención breve tiene que ver con el **conteo de datos**. Aquí a menudo es conveniente suponer una respuesta de Poisson. En aplicaciones biomédicas la respuesta que se modela en comparación con las dosis de medicamentos podría ser el número de colonias de células cancerosas. En la industria textil una respuesta razonable que se modela en comparación con ciertas variables de los procesos es el número de imperfecciones por yarda de tela.

Varianza no homogénea

El lector debería notar la comparación de la situación ideal, es decir, la respuesta normal, con la de la respuesta de Bernoulli (o binomial) o la de Poisson. Nos hemos acostumbrado al hecho de que el caso normal es muy especial debido a que la varianza es **independiente de la media**. Resulta claro que éste no es el caso para la respuesta de Bernoulli ni la de Poisson. Por ejemplo, si la respuesta es 0 o 1, lo cual sugiere una respuesta de Bernoulli, entonces el modelo adopta la forma

$$p = f(\mathbf{x}, \beta),$$

donde p es la **probabilidad de un éxito** (por ejemplo, la respuesta $= 1$). El parámetro p desempeña el papel de $\mu_{Y|x}$ en el caso normal. Sin embargo, la varianza de Bernoulli es $p(1 - p)$ que, desde luego, también es una función del regresor $\mathbf{x}$. Como resultado, la varianza no es constante. Esto descarta el uso de los mínimos cuadrados estándar que hemos utilizado en nuestro trabajo de regresión lineal hasta este momento. Lo mismo se aplica para el caso de Poisson, ya que el modelo adopta la forma

$$\lambda = f(\mathbf{x}, \beta),$$

con $\text{Var}(y) = \mu_y = \lambda$, que varía con $\mathbf{x}$.

Respuesta binaria (regresión logística)

El enfoque más popular para modelar respuestas binarias es la técnica llamada **regresión logística**, la cual se emplea mucho en las ciencias biológicas, en la investigación biomédica y en la ingeniería. De hecho, se observa que incluso en las ciencias sociales abundan las respuestas binarias. La distribución básica para la respuesta es la de Bernoulli o la binomial. La primera se encuentra en estudios observacionales donde no hay corridas repetidas en cada nivel de regresor; mientras que la segunda será el caso en que se utilice un diseño experimental. Por ejemplo, en un ensayo clínico en el cual se evalúa un fármaco nuevo, el objetivo podría ser el de determinar la dosis del medicamento que es

eficaz. Así, en el experimento se utilizarán ciertas dosis y para cada una de ellas se emplearán a varios sujetos, un caso al que se le denomina **caso agrupado**.

¿Cuál es el modelo para la regresión logística?

En el caso de respuestas binarias la respuesta media es una probabilidad. En la ilustración anterior del ensayo clínico podríamos decir que deseamos estimar la probabilidad de que el paciente responda en forma adecuada al fármaco, P(éxito). Entonces, el modelo se escribe en términos de una probabilidad. Dados los regresores $\mathbf{x}$, la función logística es dada por

$$p = \frac{1}{1 + e^{-\mathbf{x}'\beta}}.$$

La porción $\mathbf{x}'\beta$ se llama **predictor lineal** y, en el caso de un solo regresor x, se puede escribir $\mathbf{x}'\beta = \beta_0 + \beta_1 x$. Desde luego, no descartamos la inclusión de regresores múltiples y de términos polinomiales en el llamado predictor lineal. En el caso agrupado el modelo implica el modelado de la media de una binomial en vez de una de Bernoulli, por lo que la media es dada por

$$np = \frac{n}{1 + e^{-\mathbf{x}'\beta}}.$$

Características de la función logística

Una gráfica de la función logística revela mucho sobre sus características y del porqué se utiliza para este tipo de problema. En primer lugar, la función es no lineal. Además, la gráfica de la figura 12.8 revela la forma de S con la función que tiende a la asíntota en $p = 1.0$. En este caso, $\beta_1 > 0$. Así, nunca se experimentaría una probabilidad estimada mayor que 1.0.

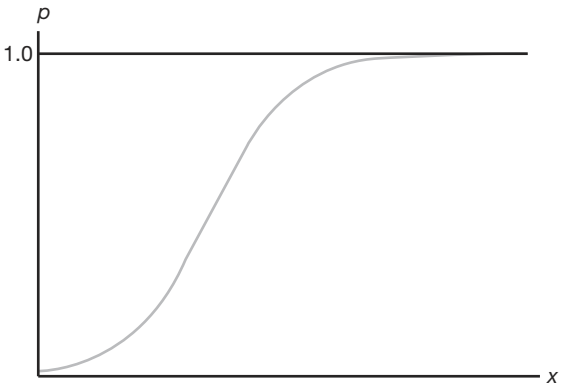


Figura 12.8: La función logística.

Los coeficientes de regresión en el predictor lineal se estiman con el método de probabilidad máxima, tal como se describió en el capítulo 9. La solución de las ecuaciones

de probabilidad requiere una metodología iterativa que no se describe aquí. Sin embargo, presentaremos un ejemplo y analizaremos la salida de resultados por computadora y las conclusiones.

Ejemplo 12.13: El conjunto de datos de la tabla 12.16 se utilizará con el fin de ilustrar el uso de la regresión logística para analizar un ensayo biológico cuantal de agente único en un experimento de toxicidad. Los resultados muestran el efecto de diferentes dosis de nicotina en la mosca común de la fruta.

Tabla 12.16: Conjunto de datos para el ejemplo 12.13

x Concentración (gramos/100 cc)	n_i Número de insectos	y Número de muertes	Porcentaje de muertes
0.10	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

El propósito del experimento era el de obtener un modelo adecuado que relacionara la probabilidad de "muerte" con la concentración. Además, el analista buscaba la denominada **dosis eficaz** (DE), es decir, la concentración de nicotina que da como resultado cierta probabilidad. La DE_{50} tiene interés particular, ya que es la concentración que produce una probabilidad de 0.5 de que el "insecto muera".

Este ejemplo es agrupado, por lo que el modelo es dado por

$$E(Y_i) = n_i p_i = \frac{n_i}{1 + e^{-(\beta_0 + \beta_1 x_i)}}.$$

Los estimados de β_0 y β_1, y sus errores estándar, se calculan usando el método de probabilidad máxima. Las pruebas de los coeficientes individuales se calculan utilizando el estadístico χ^2 en lugar del estadístico t, puesto que no hay una varianza común σ^2. El estadístico χ^2 se obtiene a partir de (coef /error estándar)^2.

Por consiguiente, obtenemos la siguiente salida de resultados por computadora de la función PROC LOGIST del *SAS*.

	Análisis de los estimados de los parámetros				
	gl	Estimado	Error estándar	Chi cuadrada	Valor P
β_0	1	−1.7361	0.2420	51.4482	< 0.0001
β_1	1	6.2954	0.7422	71.9399	< 0.0001

Ambos coeficientes difieren significativamente de cero. Por consiguiente, el modelo ajustado que se emplea para predecir la probabilidad de "muerte" es dado por

$$\hat{p} = \frac{1}{1 + e^{-(-1.7361 + 6.2954\, x)}}.$$

Estimado de la dosis eficaz

El estimado de la DE_{50} para el ejemplo 12.13 se calcula de manera muy sencilla a partir de los estimados b_0 para β_0 y b_1 para β_1. A partir de la función logística se observa que

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

Como resultado, para $p = 0.5$ se calcula un estimado de x a partir de

$$b_0 + b_1 x = 0.$$

Así, DE_{50} es dada por

$$x = -\left(\frac{b_0}{b_1}\right) = 0.276 \text{ gramos/100 cc.}$$

Concepto de razón de probabilidad

Otra forma de inferencia que se lleva a cabo de manera adecuada usando la regresión logística se deriva del uso de la razón de probabilidad, la cual está diseñada para determinar cómo se incrementa la **probabilidad de éxitos**, $\frac{p}{1-p}$, a medida que ocurren ciertos cambios en los valores del regresor. Por ejemplo, en el caso del ejemplo 12.13, quizá se deseara saber cómo aumentarían las probabilidades si la dosis se incrementara en, digamos, 0.2 gramos/100 cc.

Definición 12.1: En la regresión logística una **razón de probabilidad** es la razón de la probabilidad de éxito en la condición 2 con respecto a la de la condición 1 en los regresores, es decir,

$$\frac{[p/(1-p)]_2}{[p/(1-p)]_1}.$$

Esto permite que el analista tenga una idea de la utilidad de cambiar el regresor en cierto número de unidades. Ahora, como $\left(\frac{p}{1-p}\right) = e^{\beta_0 + \beta_1 x}$, para el ejemplo 12.13 la razón que refleja el incremento de las probabilidades de éxito cuando aumenta la dosis de nicotina en 0.2 gramos/100 cc es dada por

$$e^{0.2b_1} = e^{(0.2)(6.2954)} = 3.522.$$

La implicación de una razón de probabilidad de 3.522 es que la probabilidad de éxito aumenta en un factor de 3.522 cuando la dosis de nicotina aumenta en 0.2 gramos/100 cc.

Ejercicios

12.60 A partir de un conjunto de datos de respuestas a la dosis de estreptomicina un investigador desea desarrollar una relación entre la proporción de linfoblastos muestreados que contienen aberraciones y la dosis del medicamento. Se aplicaron cinco niveles de dosis a los conejos que se emplearon para el experimento. Los datos son los siguientes (véase Myers, 1990, listado en la bibliografía):

Dosis (mg/kg)	Número de linfoblastos	Número de aberraciones
0	600	15
30	500	96
60	600	187
75	300	100
90	300	145

a) Ajuste una regresión logística al conjunto de datos, y así estime β_0 y β_1 en el modelo

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}},$$

donde n es el número de linfoblastos, x es la dosis y p la probabilidad de una aberración.

b) Muestre los resultados de pruebas χ^2 que revelen la significancia de los coeficientes de regresión β_0 y β_1.

c) Estime la DE_{50} e interprétela.

12.61 En un experimento para estudiar el efecto de la carga, x, en lb/pulgadas^2, sobre la probabilidad de falla de especímenes de cierto tipo de tela, varios especímenes se expusieron a cargas de entre 5 lb/pulg^2 a 90 lb/pulg^2. Se observaron los números de "fallas". Los datos son los siguientes:

Carga	Número de especímenes	Número de fallas
5	600	13
35	500	95
70	600	189
80	300	95
90	300	130

a) Utilice regresión logística para ajustar el modelo

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}},$$

donde p es la probabilidad de falla y x es la carga.

b) Emplee el concepto de razón de probabilidad para determinar el incremento de la probabilidad de falla que resulta de aumentar la carga en 20 lb/pulg^2.

Ejercicios de repaso

12.62 En el Departamento de Pesca y Vida Silvestre de Virginia Tech se realizó un experimento para estudiar el efecto de las características de la corriente sobre la biomasa de los peces. Las variables regresoras son las siguientes: profundidad promedio (de 50 células), x_1; área de la cubierta en la corriente, es decir, riberas socavadas, troncos, cantos rodados, etc., x_2; porcentaje de cubierta de material translúcido (promedio de 12), x_3; y un área ≥ 25 centímetros de profundidad, x_4. La respuesta es y, la biomasa de los peces. Los datos son los siguientes:

Obs.	y	x_1	x_2	x_3	x_4
1	100	14.3	15.0	12.2	48.0
2	388	19.1	29.4	26.0	152.2
3	755	54.6	58.0	24.2	469.7
4	1288	28.8	42.6	26.1	485.9
5	230	16.1	15.9	31.6	87.6
6	0	10.0	56.4	23.3	6.9
7	551	28.5	95.1	13.0	192.9
8	345	13.8	60.6	7.5	105.8
9	0	10.7	35.2	40.3	0.0
10	348	25.9	52.0	40.3	116.6

a) Ajuste una regresión lineal múltiple que incluya las cuatro variables regresoras.

b) Utilice C_p, R^2 y s^2 para determinar el mejor subconjunto de variables. Calcule dichos estadísticos para todos los subconjuntos posibles.

c) Compare lo adecuado de los modelos de los incisos *a*) y *b*) para efectos de predecir la biomasa de los peces.

12.63 Demuestre que, en un conjunto de datos de regresión lineal múltiple,

$$\sum_{i=1}^{n} h_{ii} = p.$$

12.64 Se efectuó un experimento sencillo para ajustar una ecuación de regresión múltiple que relaciona al producto, y, con la temperatura, x_1, el tiempo de reacción, x_2, y la concentración de uno de los reactivos, x_3. Se eligieron dos niveles de cada variable y se hicieron las siguientes mediciones correspondientes a las variables independientes definidas:

y	x_1	x_2	x_3
7.6	−1	−1	−1
5.5	1	−1	−1
9.2	−1	1	−1
10.3	−1	−1	1
11.6	1	1	−1
11.1	1	−1	1
10.2	−1	1	1
14.0	1	1	1

a) Utilice las variables codificadas y estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

b) Separe la *SCR*, es decir, la suma de cuadrados de regresión, en tres componentes con un grado de libertad, atribuibles a x_1, x_2 y x_3, respectivamente. Construya una tabla de análisis de varianza donde se indiquen pruebas de significancia sobre cada variable. Comente los resultados.

12.65 En un experimento de ingeniería química relacionado con la transferencia de calor en una capa de fluido superficial se recabaron datos sobre las cuatro variables regresoras siguientes: la tasa de flujo del gas fluido en lb/hr (x_1), la tasa de flujo del gas flotante en lb/hr (x_2), la abertura de la boquilla de entrada del gas flotante en milímetros (x_3) y la temperatura de entrada del gas flotante en °F (x_4). Las respuestas medidas son la eficacia de la transferencia de calor (y_1) y la eficacia térmica (y_2). Los datos son los siguientes:

Obs.	y_1	y_2	x_1	x_2	x_3	x_4
1	41.852	38.75	69.69	170.83	45	219.74
2	155.329	51.87	113.46	230.06	25	181.22
3	99.628	53.79	113.54	228.19	65	179.06
4	49.409	53.84	118.75	117.73	65	281.30
5	72.958	49.17	119.72	117.69	25	282.20
6	107.702	47.61	168.38	173.46	45	216.14
7	97.239	64.19	169.85	169.85	45	223.88
8	105.856	52.73	169.85	170.86	45	222.80
9	99.348	51.00	170.89	173.92	80	218.84
10	111.907	47.37	171.31	173.34	25	218.12
11	100.008	43.18	171.43	171.43	45	219.20
12	175.380	71.23	171.59	263.49	45	168.62
13	117.800	49.30	171.63	171.63	45	217.58
14	217.409	50.87	171.93	170.91	10	219.92
15	41.725	54.44	173.92	71.73	45	296.60
16	151.139	47.93	221.44	217.39	65	189.14
17	220.630	42.91	222.74	221.73	25	186.08
18	131.666	66.60	228.90	114.40	25	285.80
19	80.537	64.94	231.19	113.52	65	286.34
20	152.966	43.18	236.84	167.77	45	221.72

Considere el modelo para predecir la respuesta del coeficiente de transferencia de calor

$$y_{1i} = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ji} + \sum_{i=1}^{4} \beta_{jj} x_{ji}^2 + \sum\sum_{j \neq l} \beta_{jl} x_{ji} x_{li} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

a) Calcule PRESS y $\sum_{i=1}^{n} \left| y_i - \hat{y}_{i,-i} \right|$ para ajustar el modelo anterior con los mínimos cuadrados de regresión.

b) Ajuste un modelo de segundo orden con x_4 eliminada por completo, es decir, elimine todos los términos que impliquen x_4. Calcule los criterios de predicción para el modelo reducido. Comente qué tan adecuada es x_4 para predecir el coeficiente de transferencia de calor.

c) Repita los incisos *a*) y *b*) para la eficacia térmica.

12.66 En la fisiología del deporte una medición objetiva de la condición física es el consumo de oxígeno en volumen por unidad de peso corporal por unidad de tiempo. Se estudiaron 31 individuos en un experimento con el fin de modelar el consumo de oxígeno en comparación con la edad en años, x_1, el peso en kilogramos, x_2, el tiempo para correr 1 1/2 millas, x_3, las pulsaciones en reposo, x_4, las pulsaciones al final de la carrera, x_5, y las pulsaciones máximas durante la carrera, x_6.

a) Realice una regresión por etapas a un nivel de significancia de 0.25 en la entrada. Proporcione el modelo final.

b) Estudie todos los subconjuntos posibles usando s^2, C_p, R^2 y R^2_{ajus}. Tome una decisión y determine el modelo final.

ID	y	x_1	x_2	x_3	x_4	x_5	x_6
1	44.609	44	89.47	11.37	62	178	182
2	45.313	40	75.07	10.07	62	185	185
3	54.297	44	85.84	8.65	45	156	168
4	59.571	42	68.15	8.17	40	166	172
5	49.874	38	89.02	9.22	55	178	180
6	44.811	47	77.45	11.63	58	176	176
7	45.681	40	75.98	11.95	70	176	180
8	49.091	43	81.19	10.85	64	162	170
9	39.442	44	81.42	13.08	63	174	176
10	60.055	38	81.87	8.63	48	170	186
11	50.541	44	73.03	10.13	45	168	168
12	37.388	45	87.66	14.03	56	186	192
13	44.754	45	66.45	11.12	51	176	176
14	47.273	47	79.15	10.60	47	162	164
15	51.855	54	83.12	10.33	50	166	170
16	49.156	49	81.42	8.95	44	180	185
17	40.836	51	69.63	10.95	57	168	172
18	46.672	51	77.91	10.00	48	162	168
19	46.774	48	91.63	10.25	48	162	164
20	50.388	49	73.37	10.08	76	168	168
21	39.407	57	73.37	12.63	58	174	176
22	46.080	54	79.38	11.17	62	156	165
23	45.441	52	76.32	9.63	48	164	166
24	54.625	50	70.87	8.92	48	146	155
25	45.118	51	67.25	11.08	48	172	172
26	39.203	54	91.63	12.88	44	168	172
27	45.790	51	73.71	10.47	59	186	188
28	50.545	57	59.08	9.93	49	148	155
29	48.673	49	76.32	9.40	56	186	188
30	47.920	48	61.24	11.50	52	170	176
31	47.467	52	82.78	10.50	53	170	172

12.67 Considere los datos del ejercicio de repaso 12.64. Suponga que le interesa agregar algunos términos de "interacción". En específico, considere el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \beta_{23} x_{2i} x_{3i} + \beta_{123} x_{1i} x_{2i} x_{3i} + \epsilon_i.$$

a) ¿Aún se tiene ortogonalidad? Comente al respecto.

b) Con el modelo ajustado del inciso *a*), ¿puede usted encontrar intervalos de predicción y de confianza sobre la respuesta media? Explique su respuesta.

c) Considere un modelo en el que se eliminó $\beta_{123} x_1 x_2 x_3$. Para determinar si son necesarias las interacciones (como un todo), pruebe

$$H_0: \ \beta_{12} = \beta_{13} = \beta_{23} = 0.$$

Proporcione el valor *P* y saque conclusiones.

12.68 Para extraer petróleo crudo se utiliza una técnica de inyección de dióxido de carbono (CO_2). El flujo de CO_2 envuelve el petróleo y lo desplaza. En un experimento se introducen tubos de flujo en muestras de cavidades de petróleo que contienen una cantidad conocida del mismo. Se utilizan tres valores diferentes de

presión de flujo y tres valores diferentes de ángulos de introducción, las cavidades de petróleo se inyectan con CO_2 y se registra el porcentaje de petróleo desplazado. Considere el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i.$$

Ajuste el modelo anterior a los datos y sugiera cualquier modificación al modelo que considere necesaria.

Presión lb/pulg^2, x_1	Ángulo de inyección, x_2	Recuperación de petróleo, (%), y
1000	0	60.58
1000	15	72.72
1000	30	79.99
1500	0	66.83
1500	15	80.78
1500	30	89.78
2000	0	69.18
2000	15	80.31
2000	30	91.99

Fuente: Wang, G. C. "Microscopic Investigations of CO_2 Flooding Process", Journal of Petroleum Technology, vol. 34, núm. 8, agosto de 1982.

12.69 Un artículo del *Journal of Pharmaceutical Sciences* (vol. 80, 1991) presenta datos de la solubilidad de una fracción molar de un soluto a temperatura constante. También se midió la dispersión, x_1, y los parámetros de solubilidad del enlace bipolar y de hidrógeno, x_2 y x_3. En la tabla siguiente se presenta una parte de los datos. En el modelo, y es el logaritmo negativo de la fracción molar. Ajuste el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i,$$

para $i = 1, 2,..., 20$.

Obs.	y	x_1	x_2	x_3
1	0.2220	7.3	0.0	0.0
2	0.3950	8.7	0.0	0.3
3	0.4220	8.8	0.7	1.0
4	0.4370	8.1	4.0	0.2
5	0.4280	9.0	0.5	1.0
6	0.4670	8.7	1.5	2.8
7	0.4440	9.3	2.1	1.0
8	0.3780	7.6	5.1	3.4
9	0.4940	10.0	0.0	0.3
10	0.4560	8.4	3.7	4.1
11	0.4520	9.3	3.6	2.0
12	0.1120	7.7	2.8	7.1
13	0.4320	9.8	4.2	2.0
14	0.1010	7.3	2.5	6.8
15	0.2320	8.5	2.0	6.6
16	0.3060	9.5	2.5	5.0
17	0.0923	7.4	2.8	7.8
18	0.1160	7.8	2.8	7.7
19	0.0764	7.7	3.0	8.0
20	0.4390	10.3	1.7	4.2

a) Pruebe H_0: $\beta_1 = \beta_2 = \beta_3 = 0$.
b) Grafique los residuales estudentizados en comparación con x_1, x_2 y x_3 (tres gráficas). Haga comentarios al respecto.
c) Considere dos modelos adicionales que compitan con el modelo anterior:

Modelo 2: Agregue x_1^2, x_2^2, x_3^2.

Modelo 3: Agregue $x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$.

Utilice PRESS y C_p con estos tres modelos para saber cuál de los tres es el mejor.

12.70 Se realizó un estudio para determinar si los cambios en el estilo de vida podrían sustituir la medicación para reducir la presión sanguínea de los individuos hipertensos. Los factores considerados fueron una dieta saludable con un programa de ejercicios, la dosis común de medicamentos para la hipertensión y ningún tratamiento. También se calculó el índice de masa corporal (IMC) previo al tratamiento, debido a que se sabe que éste afecta la presión sanguínea. La respuesta considerada en este estudio fue el cambio en la presión sanguínea. La variable "grupo" tenía los siguientes niveles.

1 = Dieta saludable y programa de ejercicios

2 = Medicación

3 = Sin tratamiento

a) Ajuste un modelo adecuado utilizando los datos anteriores. ¿Parece que el ejercicio y la dieta se pueden utilizar en forma eficaz para disminuir la presión sanguínea? Explique su respuesta a partir de los resultados.
b) ¿El ejercicio y la dieta son una alternativa eficaz a la medicación?

(*Sugerencia*: Para responder a estas preguntas quizás usted desee construir el modelo en más de una forma).

Cambio en la presión sanguínea	Grupo	IMC
−32	1	27.3
−21	1	22.1
−26	1	26.1
−16	1	27.8
−11	2	19.2
−19	2	26.1
−23	2	28.6
−5	2	23.0
−6	3	28.1
5	3	25.3
−11	3	26.7
14	3	22.3

12.71 Demuestre que al elegir el llamado mejor modelo del subconjunto de entre una serie de posibles modelos, elegir el modelo con la menor s^2 equivale a escoger el modelo con el R^2_{ajus} más pequeño.

12.72 Estudio de caso: Considere el conjunto de datos para el ejercicio 12.12 de la página 452 (datos de un hospital) que se repite a continuación.

Sitio	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1003.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11,520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20,106	3655.08	180.5	6.15	3503.93
11	96.00	13,313	2912.00	60.9	5.88	3571.59
12	131.42	10,771	3921.00	103.7	4.88	3741.40
13	127.21	15,543	3865.67	126.8	5.50	4026.52
14	252.90	36,194	7684.10	157.7	7.00	10,343.81
15	409.20	34,703	12,446.33	169.4	10.75	11,732.17
16	463.70	39,204	14,098.40	331.4	7.05	15,414.94
17	510.22	86,533	15,524.00	371.6	6.35	18,854.45

a) Los listados de resultados de la función PROC REG del *SAS* que se presentan en las figuras 12.9 y 12.10 proporcionan una cantidad considerable de información. El propósito es detectar los valores extremos y, a final de cuentas, determinar cuáles términos del modelo deben utilizarse en la versión final.

b) A menudo ocurre que el papel que desempeña una sola variable regresora no es evidente cuando se estudia en presencia de otras variables; esto se debe a la multicolinealidad. Con esto presente haga comentarios sobre la importancia de x_2 y x_3 en el modelo completo en comparación con su importancia en un modelo en el cual éstas son las únicas variables.

c) Comente acerca de qué otros análisis se tendrían que hacer.

d) Elabore análisis apropiados y escriba sus conclusiones respecto al modelo final.

```
Dependent Variable: y
                      Analysis of Variance
                                   Sum of           Mean
Source                 DF          Squares         Square      F Value     Pr > F
Model                   5        490177488       98035498       237.79     <.0001
Error                  11          4535052         412277
Corrected Total        16        494712540

              Root MSE            642.08838    R-Square     0.9908
              Dependent Mean     4978.48000    Adj R-Sq     0.9867
              Coeff Var            12.89728

                                   Parameter Estimates
                                             Parameter  Standard
Variable  Label                          DF    Estimate      Error t Value  Pr > |t|
Intercept Intercept                       1  1962.94816 1071.36170    1.83    0.0941
x1        Average Daily Patient Load      1   -15.85167   97.65299   -0.16    0.8740
x2        Monthly X-Ray Exposure          1     0.05593    0.02126    2.63    0.0234
x3        Monthly Occupied Bed Days       1     1.58962    3.09208    0.51    0.6174
x4        Eligible Population in the      1    -4.21867    7.17656   -0.59    0.5685
          Area/100
x5        Average Length of Patients      1   -394.31412 209.63954   -1.88    0.0867
          Stay in Days
```

Figura 12.9: Salida de resultados del *SAS* para el ejercicio de repaso 12.72; parte I.

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	566.5200	775.0251	241.2323	244.0765	1306	-734.6494	2285
2	696.8200	740.6702	331.1402	11.8355	1470	-849.4275	2331
3	1033	1104	278.5116	490.9234	1717	-436.5244	2644
4	1604	1240	268.1298	650.3459	1831	-291.0028	2772
5	1611	1564	211.2372	1099	2029	76.6816	3052
6	1613	2151	279.9293	1535	2767	609.5796	3693
7	1854	1690	218.9976	1208	2172	196.5345	3183
8	2161	1736	468.9903	703.9948	2768	-13.8306	3486
9	2306	2737	290.4749	2098	3376	1186	4288
10	3504	3682	585.2517	2394	4970	1770	5594
11	3572	3239	189.0989	2823	3655	1766	4713
12	3741	4353	328.8507	3630	5077	2766	5941
13	4027	4257	314.0481	3566	4948	2684	5830
14	10344	8768	252.2617	8213	9323	7249	10286
15	11732	12237	573.9168	10974	13500	10342	14133
16	15415	15038	585.7046	13749	16328	13126	16951
17	18854	19321	599.9780	18000	20641	17387	21255

Obs	Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	-208.5051	595.0	-0.350	| | |
2	-43.8502	550.1	-0.0797	| | |
3	-70.7734	578.5	-0.122	| | |
4	363.1244	583.4	0.622	| |* |
5	46.9483	606.3	0.0774	| | |
6	-538.0017	577.9	-0.931	|* | |
7	164.4696	603.6	0.272	| | |
8	424.3145	438.5	0.968	| |* |
9	-431.4090	572.6	-0.753	|* | |
10	-177.9234	264.1	-0.674	|* | |
11	332.6011	613.6	0.542	| |* |
12	-611.9330	551.5	-1.110	| ** | |
13	-230.5684	560.0	-0.412	| | |
14	1576	590.5	2.669	| |***** |
15	-504.8574	287.9	-1.753	| *** | |
16	376.5491	263.1	1.431	| |** |
17	-466.2470	228.7	-2.039	| **** | |

Figura 12.10: Salida de resultados del *SAS* para el ejercicio de repaso 12.72; parte II.

12.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

En este capítulo estudiamos varios procedimientos para usarlos en un "intento" por encontrar el mejor modelo. Sin embargo, una de las confusiones más importantes en el trabajo de los científicos e ingenieros novatos es que existe un **modelo lineal verdadero**, y que es posible encontrarlo. En la mayoría de fenómenos de la ciencia las relaciones entre las variables científicas son de naturaleza no lineal y se desconoce el modelo verdadero. Los modelos estadísticos lineales son **aproximaciones empíricas**.

En ocasiones, la decisión sobre cuál modelo adoptar depende de la información que se necesita obtener de éste. ¿Se usará para realizar predicciones? ¿Para explicar el papel que desempeña cada regresor? Esta "decisión" podría ser difícil ante la presencia de colinealidad. Es un hecho que para muchos problemas de regresión hay modelos múltiples con un desempeño muy similar. Para mayores detalles véase la referencia de Myers (1990).

Uno de los abusos más nocivos del material de este capítulo consiste en dar demasiada importancia a R^2 en la selección del llamado mejor modelo. Es importante recordar que para cualquier conjunto de datos se puede obtener una R^2 tan grande como se desee, dentro de la restricción de que $0 \leq R^2 \leq 1$. **Prestar demasiada atención a R^2 con frecuencia conduce a un sobreajuste**.

En este capítulo se dio mucha importancia a la detección de los valores extremos. Un clásico y grave abuso de la estadística radica en la decisión relacionada con la detección de los valores extremos. Esperamos que quede claro que el analista no debería por ningún motivo detectar los valores extremos, eliminarlos del conjunto de datos, ajustar un modelo nuevo, informar sobre los valores extremos, y así sucesivamente. Se trata de un procedimiento tentador y desastroso para llegar a un modelo que se ajuste bien a los datos, el cual conlleva a un ejemplo de **cómo mentir con estadísticos**. Si se detecta un valor extremo, lo correcto es revisar la historia de los datos en busca de posibles errores de captura o de procedimiento antes de eliminarlos del conjunto de datos. Se debe recordar que, por definición, un valor extremo es aquel para el cual el modelo no se ajusta bien. El problema podría no estar en los datos sino en la selección del modelo. Cambiar el modelo quizás haría que el punto no se detecte como un valor extremo.

Existen muchos tipos de respuestas que ocurren de forma natural en la práctica, pero que no se pueden utilizar en un análisis de mínimos cuadrados estándar porque sus supuestos de mínimos cuadrados clásicos no se cumplen. Los supuestos que suelen fallar son los de los errores normales y de la varianza homogénea. Por ejemplo, si la respuesta es una proporción, digamos la proporción de artículos defectuosos, la distribución de las respuestas se relaciona con la distribución binomial. Una segunda respuesta que ocurre con frecuencia en la práctica es la del conteo de Poisson. Evidentemente, la distribución no es normal, y la varianza de la respuesta, que es igual a la media de Poisson, varía de una observación a otra. Para conocer más detalles sobre estas condiciones poco ideales véase Myers y colaboradores (2008), citado en la bibliografía.

Capítulo 13

Experimentos con un solo factor: generales

13.1 Técnica del análisis de varianza

En el material sobre estimación y prueba de hipótesis que se cubrió en los capítulos 9 y 10 en cada caso nos limitamos a considerar sólo dos parámetros de la población. Ése fue el caso, por ejemplo, en la prueba de la igualdad de dos medias de la población, en la cual se usaron muestras independientes de poblaciones normales con varianza común pero desconocida, y en donde se necesitaba obtener un estimado agrupado de σ^2.

El material que se refiere a las inferencias de dos muestras representa un caso especial de lo que se denomina *problema de un solo factor*. Por ejemplo, en el ejercicio 10.35 de la página 357 se midió el tiempo de supervivencia para dos muestras de ratones, en donde una muestra recibió un tratamiento de suero contra la leucemia y la otra no lo recibió. En este caso decimos que hay *un factor*, es decir, un *tratamiento*, y el factor se halla en *dos niveles*. Si en el proceso de muestreo se utilizaran varios tratamientos en competencia, se necesitarían más muestras de ratones. En ese caso el problema implicaría un factor con más de dos niveles, por lo tanto, con más de dos muestras.

En el problema de $k > 2$ muestras se supone que hay k muestras provenientes de k poblaciones. Un procedimiento muy común que se utiliza cuando se prueban medias de la población se denomina **análisis de varianza**, o **ANOVA**.

Si el lector ha estudiado el material acerca de la teoría de la regresión, el análisis de varianza no será, por supuesto, una técnica nueva para él. Utilizamos el método del análisis de varianza para partir la suma total de cuadrados en dos partes, una parte debida a la regresión y otra debida al error.

Suponga que en un experimento industrial a un ingeniero le interesa la forma en que la absorción media de humedad del concreto varía para 5 agregados de concreto diferentes. Las muestras se exponen a la humedad durante 48 horas y se decide que para cada agregado deben probarse 6 muestras, lo que hace que se requiera probar un total de 30 muestras. En la tabla 13.1 se presentan los datos registrados.

El modelo que se considera para esta situación es el siguiente. Se tomaron 6 observaciones de cada una de las 5 poblaciones, con medias μ_1, μ_2,..., μ_5, respectivamente. Deseamos probar

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_5,$$

H_1: Al menos dos de las medias no son iguales.

Tabla 13.1: Absorción de humedad en agregados para concreto

Agregado:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Media	553.33	569.33	610.50	465.17	610.67	561.80

Además, estamos interesados en realizar comparaciones individuales entre estas 5 medias de la población.

Dos fuentes de variabilidad en los datos

En el procedimiento del análisis de varianza se supone que cualquier variación que exista entre los promedios de los agregados se atribuye a 1) la variación en la absorción entre observaciones *dentro* de los tipos de agregados, y 2) la variación *entre* los tipos de agregados, es decir, a las diferencias en la composición química de los agregados. Por supuesto, la **variación dentro de los agregados** se debe a varias causas. Quizá las condiciones de temperatura y humedad no se mantuvieron constantes durante el experimento. Es posible que haya habido cierta cantidad de heterogeneidad en los lotes de materias primas que se usaron. En todo caso debe considerarse la variación dentro de la muestra como una **variación aleatoria o al azar**. Parte del objetivo del análisis de varianza consiste en determinar si las diferencias entre las 5 medias muestrales son lo que se esperaría debido sólo a la variación aleatoria o si, más bien, se trata de una variación más allá de los simples efectos del azar, como las diferencias en la composición química de los agregados.

En esta etapa surgen muchas preguntas acerca del problema anterior. Por ejemplo, ¿cuántas muestras deben probarse para cada agregado? Ésta es una pregunta que desafía continuamente al analista. Además, ¿qué pasaría si la variación dentro de la muestra fuera tan grande que al procedimiento estadístico le resultara difícil detectar las diferencias sistemáticas? ¿Es posible controlar de manera sistemática fuentes externas de variación y así eliminarlas de la parte que llamamos variación aleatoria? En las secciones siguientes intentaremos responder éstas y otras preguntas.

13.2 La estrategia del diseño de experimentos

En los capítulos 9 y 10 se estudiaron los conceptos de la estimación y la prueba de hipótesis para el caso de dos muestras, bajo la importante perspectiva de la manera en que se realiza el experimento. Esto forma parte de la categoría amplia de los diseños experimentales. Por ejemplo, para la **prueba *t* agrupada** que se estudió en el capítulo 10, se supone que los niveles de los factores (los tratamientos en el ejemplo de los ratones) se asignan al azar a las unidades experimentales (los ratones). En los capítulos 9 y 10 analizamos el

concepto de unidades experimentales y lo ilustramos por medio de varios ejemplos. En pocas palabras, las unidades experimentales son las unidades (ratones, pacientes, especímenes de concreto, tiempo) que **proporcionan la heterogeneidad que conduce al error experimental** en una investigación científica. La asignación aleatoria elimina el sesgo que podría originarse con una asignación sistemática. El objetivo consiste en distribuir en forma uniforme entre los niveles de los factores los riesgos que introduce la heterogeneidad de las unidades experimentales. Una asignación al azar simula mejor las condiciones que se asumen en el modelo. En la sección 13.7 analizamos la **formación de bloques** en los experimentos. En los capítulos 9 y 10 se presentó el concepto, cuando se efectuaron comparaciones entre las medias usando el **emparejamiento**, es decir, la división de las unidades experimentales en pares homogéneos denominados **bloques**. Entonces, los niveles de los factores o tratamientos se asignan al azar dentro de los bloques. El propósito de la formación de bloques es reducir el error experimental efectivo. En este capítulo se extiende de manera natural el emparejamiento a bloques de tamaño mayor, con el análisis de varianza como la herramienta analítica principal.

13.3 Análisis de varianza de un factor: diseño completamente aleatorizado (ANOVA de un factor)

De k poblaciones se seleccionan muestras aleatorias de tamaño n. Las k poblaciones diferentes se clasifican con base en un criterio único, como tratamientos o grupos distintos. En la actualidad el término **tratamiento** se utiliza por lo general para designar las diversas clasificaciones, ya sean diferentes agregados, analistas, fertilizadores o regiones del país.

Suposiciones e hipótesis del ANOVA de un solo factor

Se supone que las k poblaciones son independientes y que están distribuidas en forma normal con medias $\mu_1, \mu_2, ..., \mu_k$, y varianza común σ^2. Como se indicó en la sección 13.2, estas suposiciones son más aceptables mediante la aleatoriedad. Se desean obtener métodos adecuados para probar las hipótesis

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k,$$

H_1: Al menos dos de las medias no son iguales.

Sea que y_{ij} denote la j-ésima observación del i-ésimo tratamiento, y el acomodo de los datos es el que se observa en la tabla 13.2. Aquí, Y_i es el total de todas las observaciones de la muestra, del i-ésimo tratamiento, $\bar{y}_i$, es la media de todas las observaciones en la muestra del i-ésimo tratamiento, $Y..$ es el total de todas las nk observaciones, y $\bar{y}...$ es la media de todas las nk observaciones.

Modelo de ANOVA para un solo factor

Cada observación puede escribirse en la forma

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

donde ϵ_{ij} mide la desviación que tiene la observación j-ésima de la i-ésima muestra, con respecto de la media del tratamiento correspondiente. El término ϵ_{ij} representa el error aleatorio y desempeña el mismo papel que los términos del error en los modelos de

Tabla 13.2: k muestras aleatorias

Tratamiento:	**1**	**2**	$\cdots$	i	$\cdots$	k	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Media	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

regresión. Una forma alternativa y preferible de esta ecuación se obtiene sustituyendo $\mu_i = \mu + \alpha_i$, sujeta a la restricción $\sum_{i=1}^{k} \alpha_i = 0$. Por lo tanto, se escribe

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

donde μ tan sólo es la **media general** de todas las μ_i, es decir,

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i,$$

y α_i se denomina el **efecto** del i-ésimo tratamiento.

La hipótesis nula de que k medias de la población son iguales, en comparación con la alternativa de que al menos dos de las medias son distintas, ahora se puede reemplazar por las hipótesis equivalentes.

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0,$
H_1: Al menos una de las α_i no es igual a cero.

Resolución de la variabilidad total en componentes

Nuestra prueba se basará en una comparación de dos estimados independientes de la varianza poblacional común σ^2. Dichos estimadores se obtendrán haciendo la partición de la variabilidad total de nuestros datos, denotados mediante la sumatoria doble

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2,$$

en dos componentes.

Teorema 13.1: Identidad de la suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

En lo que sigue, será conveniente identificar los términos de la identidad de la suma de cuadrados con la siguiente notación:

Tres medidas importantes de variabilidad

$$STC = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{suma total de cuadrados,}$$

$$SCT = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{suma de los cuadrados del tratamiento,}$$

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{suma de los cuadrados del error.}$$

Entonces, la identidad de la suma de los cuadrados se puede representar simbólicamente con la ecuación

$$STC = SCT + SCE$$

La identidad anterior expresa cómo las variaciones entre los tratamientos y dentro de los tratamientos contribuyen a la suma total de cuadrados. Sin embargo, se puede obtener mucha información si se investiga el **valor esperado tanto de *SCT* como de *SCE***. Eventualmente calcularemos estimados de la varianza que determinan la razón que

Teorema 13.2:

$$E(SCT) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2$$

La prueba del teorema se deja como ejercicio para el lector (véase el ejercicio 13.53 de la página 556).

Si H_0 es verdadera, un estimado de σ^2 basado en $k-1$ grados de libertad es dado por la expresión:

Media cuadrática del tratamiento

$$s_1^2 = \frac{SCT}{k-1}$$

Si H_0 es verdadera y por ello cada α_i en el teorema 13.2 es igual a cero, se observa que

$$E\left(\frac{SCT}{k-1}\right) = \sigma^2,$$

y s_1^2 es un estimado no sesgado de σ^2. Sin embargo, si H_1 es verdadera, se tiene que

$$E\left(\frac{SCT}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

y s_1^2 estima a σ^2 más un término adicional, que mide la variación debida a los efectos sistemáticos.

Otro estimado independiente de σ^2, basado en $k(n-1)$ grados de libertad, es la fórmula familiar:

Cuadrado medio del error

$$s^2 = \frac{SCE}{k(n-1)}$$

Resulta aleccionador puntualizar la importancia de los valores esperados de los cuadrados medios a los que recién nos referimos. En la sección siguiente se estudia el empleo de una **razón *F*** con el cuadrado medio del tratamiento en el numerador. Se observa que cuando H_1 es verdadera, la presencia de la condición $E(s_1^2) > E(s^2)$ sugiere que la razón F se utiliza en el contexto de una **prueba unilateral de cola superior**. Es decir, cuando H_1 es verdadera se esperaría que el numerador s_1^2 fuera mayor que el denominador.

Uso de la prueba *F* en el ANOVA

El estimado s^2 es no sesgado, independientemente de la veracidad o falsedad de la hipótesis nula (véase el ejercicio de repaso 13.52 de la página 556). Es importante señalar que la identidad de la suma de cuadrados ha hecho la partición no sólo de la variabilidad total de los datos, sino también del número total de grados de libertad. Es decir,

$$nk - 1 = k - 1 + k(n - 1).$$

Razón *F* para probar la igualdad de las medias

Cuando H_0 es verdadera, la razón $f = s_1^2/s^2$ es un valor de la variable aleatoria F, que tiene la distribución F con $k - 1$ y $k(n - 1)$ grados de libertad (véase el teorema 8.8). Como s_1^2 sobrestima a σ^2 cuando H_0 es falsa, se tiene una prueba de una cola con la región crítica localizada por completo en la cola derecha de la distribución.

A un nivel de significancia de α se rechaza la hipótesis nula H_0 cuando

$$f > f_\alpha[k - 1, k(n - 1)].$$

Otro método, el del valor P, sugiere que la evidencia a favor o en contra de H_0 es

$$P = P\{f[k - 1, k(n - 1)] > f\}.$$

Los cálculos para un problema de análisis de varianza por lo general se resumen en forma tabular, como se observa en la tabla 13.3.

Tabla 13.3: Análisis de varianza para el ANOVA de un solo factor

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*f* calculada
Tratamientos	SCT	$k - 1$	$s_1^2 = \frac{SCT}{k - 1}$	$\frac{s_1^2}{s^2}$
Error	SCE	$k(n - 1)$	$s^2 = \frac{SCE}{k(n - 1)}$	
Total	STC	$kn - 1$		

Ejemplo 13.1: Pruebe la hipótesis de que $\mu_1 = \mu_2 = ... = \mu_5$ a un nivel de significancia de 0.05 para los datos de la tabla 13.1 sobre la absorción de humedad por varios tipos de agregados para cemento.

***Solución*:** Las hipótesis son

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_5,$$

H_1: Al menos dos de las medias no son iguales.
$\alpha = 0.05$

Región crítica: $f > 2.76$ con $v_1 = 4$ y $v_2 = 25$ grados de libertad. Los cálculos de la suma de cuadrados proporcionan

$$STC = 209{,}377, \quad SCT = 85{,}356,$$
$$SCE = 209{,}377 - 85{,}356 = 124{,}021.$$

En la figura 13.1 se presentan estos resultados y el resto de los cálculos del procedimiento ANOVA del programa *SAS*.

```
                          The GLM Procedure
Dependent  Variable:  moisture

                                          Sum of
Source              DF          Squares      Mean Square     F Value     Pr > F
Model                4       85356.4667       21339.1167        4.30     0.0088
Error               25      124020.3333        4960.8133
Corrected  Total     29      209376.8000

  R-Square      Coeff  Var       Root  MSE     moisture  Mean
  0.407669       12.53703       70.43304           561.8000

Source              DF        Type  I  S S     Mean  Square     F Value     Pr > F
aggregate            4      85356.46667       21339.11667        4.30     0.0088
```

Figura 13.1: Salida de resultados del programa *SAS* para el procedimiento de análisis de varianza.

Decisión: Rechazar H_0 y concluir que los agregados no tienen la misma media de absorción. El valor P para $f = 4.30$ es 0.0088, que es menor que 0.05. ∎

Además del ANOVA, se construyeron gráficas de caja para cada agregado, las cuales se presentan en la figura 13.2. Al observar las gráficas vemos que es evidente que no todos los agregados tienen la misma absorción. De hecho, parece que el agregado 4 destaca del resto. En el ejercicio 13.21 de la página 531 se incluye un análisis más formal que revela este resultado.

Durante el trabajo experimental es frecuente que se pierdan algunas de las observaciones deseadas. Los animales experimentales mueren, el material experimental se daña o los seres humanos abandonan el estudio. El análisis anterior para el mismo tamaño de la muestra aún es válido si modificamos ligeramente las fórmulas de la suma de cuadrados. Ahora suponemos que las k muestras aleatorias son de tamaño $n_1, n_2, \ldots, n_k$, respectivamente.

Suma de cuadrados; tamaños desiguales de las muestras

$$STC = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2, \ SCT = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2, \ SCE = STC - SCT$$

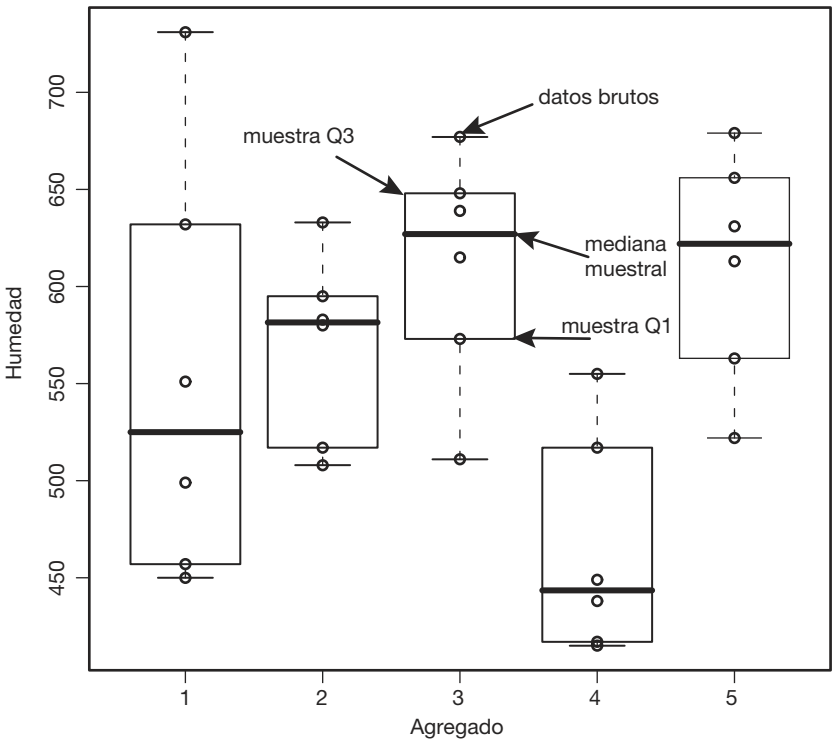


Figura 13.2: Gráficas de caja para la absorción de la humedad en agregados de concreto.

Después se hace la partición de los grados de libertad, como antes: $N-1$ para STC, $k-1$ para SCT y $N-1-(k-1)=N-k$ para SCE, donde $N=\sum_{i=1}^{k} n_i$.

Ejemplo 13.2: Parte de un estudio realizado en Virginia Tech se diseñó para medir los niveles de actividad de la fosfatasa alcalina sérica (en unidades de Bessey-Lowry) en niños con trastornos convulsivos que recibían terapia de anticonvulsivantes bajo el cuidado de un médico privado. Se reclutaron 45 sujetos para el estudio y se clasificaron en cuatro grupos de medicamentos:

G-1: Control (no recibieron anticonvulsivantes ni tenían historia de trastornos convulsivos)
G-2: Fenobarbital
G-3: Carbamazepina
G-4: Otros anticonvulsivantes

De las muestras de sangre tomadas a cada sujeto se determinó el nivel de actividad de la fosfatasa alcalina sérica y se registró tal como se observa en la tabla 13.4. Pruebe la hipótesis de que, a un nivel de significancia de 0.05, el nivel promedio de actividad de la fosfatasa alcalina sérica es el mismo para los cuatro grupos de medicamentos.

Tabla 13.4: Nivel de actividad de la fosfatasa alcalina sérica

G-1		G-2	G-3	G-4
49.20	97.50	97.07	62.10	110.60
44.54	105.00	73.40	94.95	57.10
45.80	58.05	68.50	142.50	117.60
95.84	86.60	91.85	53.00	77.71
30.10	58.35	106.60	175.00	150.00
36.50	72.80	0.57	79.50	82.90
82.30	116.70	0.79	29.50	111.50
87.85	45.15	0.77	78.40	
105.00	70.35	0.81	127.50	
95.22	77.40			

***Solución*:** A un nivel de significancia de 0.05, las hipótesis son

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$,
H_1: Al menos dos de las medias no son iguales.

Región crítica: $f > 2.836$, al interpolar los valores de la tabla A.6.
Cálculos: $Y_{1.} = 1460.25$, $Y_{2.} = 440.36$, $Y_{3.} = 842.45$, $Y_{4.} = 707.41$ y $Y_{..} = 3450.47$. El análisis de varianza se incluye en la salida de resultados de *MINITAB* que se presenta en la figura 13.3.

```
One-way ANOVA: G-1, G-2, G-3, G-4

Source   DF      SS     MS      F      P
Factor    3   13939   4646   3.57  0.022
Error    41   53376   1302
Total    44   67315

S = 36.08    R-Sq = 20.71%    R-Sq(adj ) = 14.90%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level    N     Mean   StDev  --+---------+---------+---------+-------
G-1     20    73.01   25.75              (----*-----)
G-2      9    48.93   47.11  (-------*-------)
G-3      9    93.61   46.57                   (-------*-------)
G-4      7   101.06   30.76                     (--------*--------)
                          --+---------+---------+---------+-------
                          30         60         90        120

Pooled StDev = 36.08
```

Figura 13.3: Análisis de *MINITAB* para los datos de la tabla 13.4.

Decisión: Rechazar H_0 y concluir que los niveles de actividad promedio de la fosfatasa alcalina sérica para los cuatro grupos de medicamentos no son los mismos. El valor calculado de P es 0.022. ∎

Para concluir nuestro estudio del análisis de varianza para la clasificación de un solo factor mencionaremos las ventajas de elegir muestras del mismo tamaño en vez de otras de tamaños distintos. La primera ventaja es que la razón f no es sensible a pequeñas desviaciones de la suposición de varianzas iguales para las k poblaciones cuando las muestras son del mismo tamaño. La segunda consiste en que muestras del mismo tamaño minimizan la probabilidad de cometer un error tipo II.

13.4 Pruebas de la igualdad de varias varianzas

Aunque la razón f que se obtiene con el procedimiento del análisis de varianza no es sensible a las desviaciones de la suposición de varianzas iguales para las k poblaciones normales cuando las muestras son de igual tamaño, debe tenerse precaución y efectuar una prueba preliminar sobre la homogeneidad de las varianzas. En el caso de muestras de tamaños distintos, salta a la vista que es aconsejable realizar una prueba como ésa, si existe duda razonable acerca de la homogeneidad de las varianzas de la población. Por lo tanto, suponga que se desea probar la hipótesis nula

$$H_0: \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

en comparación con la alternativa

H_1: No todas las varianzas son iguales.

La prueba que usaremos, denominada **prueba de Bartlett**, se basa en un estadístico cuya distribución muestral proporciona valores críticos exactos cuando los tamaños de las muestras son iguales. Dichos valores críticos para tamaños de las muestras iguales también se pueden utilizar para obtener aproximaciones muy exactas de los valores críticos para tamaños muestrales distintos.

En primer lugar calculamos las k varianzas muestrales s_1^2, s_2^2,..., s_k^2 a partir de muestras de tamaño n_1, n_2,..., n_k, con $\sum_{i=1}^{k} n_i = N$. En segundo lugar combinamos las varianzas muestrales para obtener la estimación agrupada

$$s_p^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) s_i^2.$$

Ahora,

$$b = \frac{[(s_1^2)^{n_1 - 1} (s_2^2)^{n_2 - 1} \cdots (s_k^2)^{n_k - 1}]^{1/(N-k)}}{s_p^2}$$

es un valor de una variable aleatoria B que tiene la **distribución de Bartlett**. Para el caso especial en que $n_1 = n_2 = \cdots = n_k = n$, se rechaza H_0 a un nivel de significancia α si

$$b < b_k(\alpha; n),$$

donde $b_k(\alpha; n)$ es el valor crítico que deja una área de tamaño α en el extremo izquierdo de la distribución de Bartlett. En la tabla A.10 se incluyen los valores críticos, $b_k(\alpha; n)$, para $\alpha = 0.01$ y 0.05; $k = 2, 3, ..., 10$; y valores seleccionados de n, desde 3 hasta 100.

Cuando los tamaños de las muestras son distintos, se rechaza la hipótesis nula al nivel de significancia α si

$$b < b_k(\alpha; n_1, n_2, \ldots, n_k),$$

donde

$$b_k(\alpha; n_1, n_2, \ldots, n_k) \approx \frac{n_1 b_k(\alpha; n_1) + n_2 b_k(\alpha; n_2) + \cdots + n_k b_k(\alpha; n_k)}{N}.$$

Igual que antes, todas las $b_k(\alpha; n_i)$ para los tamaños muestrales n_1, n_2,..., n_k se obtienen de la tabla A.10.

Ejemplo 13.3: Utilice la prueba de Bartlett a un nivel de significancia de 0.01 para probar la hipótesis de que las varianzas de la población de los cuatro grupos de medicamentos del ejemplo 13.2 son iguales.

***Solución*:** Tenemos la hipótesis

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$,
H_1: Las varianzas no son iguales,
con $\alpha = 0.01$.

Región crítica: Si nos remitimos al ejemplo 13.2, tenemos que $n_1 = 20$, $n_2 = 9$, $n_3 = 9$, $n_4 = 7$, $N = 45$ y $k = 4$. Por lo tanto, se rechaza cuando

$$\begin{aligned} b &< b_4(0.01; 20, 9, 9, 7) \\ &\approx \frac{(20)(0.8586) + (9)(0.6892) + (9)(0.6892) + (7)(0.6045)}{45} \\ &= 0.7513. \end{aligned}$$

Cálculos: El primero se obtiene

$$s_1^2 = 662.862,\ s_2^2 = 2219.781,\ s_3^2 = 2168.434,\ s_4^2 = 946.032,$$

y después

$$\begin{aligned} s_p^2 &= \frac{(19)(662.862) + (8)(2219.781) + (8)(2168.434) + (6)(946.032)}{41} \\ &= 1301.861. \end{aligned}$$

Ahora,

$$b = \frac{[(662.862)^{19}\,(2219.781)^{8}\,(2168.434)^{8}\,(946.032)^{6}]^{1/41}}{1301.861} = 0.8557.$$

Decisión: no rechazar la hipótesis y concluir que las varianzas de la población de los cuatro grupos de medicamentos no son significativamente distintas.

Aunque la prueba de Bartlett se utiliza con mayor frecuencia para probar la homogeneidad de varianzas, se dispone de otros métodos. Un método creado por Cochran proporciona un procedimiento de cálculo sencillo, aunque está limitado a situaciones en

que los tamaños muestrales son iguales. La **prueba de Cochran** es especialmente útil para detectar si alguna de las varianzas es mucho mayor que las demás. El estadístico que se emplea es:

$$G = \frac{S_i^2 \text{ más grande}}{\sum_{i=1}^{k} S_i^2},$$

y se rechaza la hipótesis de igualdad de varianzas si $g > g_\alpha$, donde el valor de g_α se obtiene de la tabla A.11.

Para ilustrar la prueba de Cochran nos remitiremos otra vez a los datos de la tabla 13.1 sobre la absorción de humedad de los agregados para concreto. ¿Se justificó la suposición de varianzas iguales al realizar el análisis de varianza en el ejemplo 13.1? Se encontró que

$$s_1^2 = 12{,}134,\ s_2^2 = 2303,\ s_3^2 = 3594,\ s_4^2 = 3319,\ s_5^2 = 3455.$$

Por lo tanto,

$$g = \frac{12{,}134}{24{,}805} = 0.4892,$$

que no excede el valor de la tabla $g_{0.05} = 0.5065$. En consecuencia, se concluye que es razonable la suposición de que las varianzas son iguales.

Ejercicios

13.1 Se están considerando seis máquinas diferentes para la fabricación de sellos de goma y se están comparando con respecto a la resistencia a la tensión del producto. Se utiliza una muestra aleatoria de cuatro sellos hechos con cada máquina para determinar si la resistencia media a la tensión varía de una máquina a otra. A continuación se presentan las medidas de la resistencia a la tensión en kilogramos por centímetro cuadrado $\times 10^{-1}$:

Máquina					
1	**2**	**3**	**4**	**5**	**6**
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Realice el análisis de varianza a un nivel de significancia de 0.05 e indique si la resistencia promedio a la tensión de las seis máquinas difiere o no de manera significativa.

13.2 Los datos que se presentan en la siguiente tabla representan el número de horas de alivio proporcionadas por cinco marcas diferentes de tabletas para el dolor de cabeza administradas a 25 sujetos que tenían fiebre de 38°C o más. Realice el análisis de varianza y, a un nivel de significancia de 0.05, pruebe la hipótesis de que las cinco marcas proporcionan el mismo número medio de horas de alivio. Analice los resultados.

Tabletas				
A	***B***	***C***	***D***	***E***
5.2	9.1	3.2	2.4	7.1
4.7	7.1	5.8	3.4	6.6
8.1	8.2	2.2	4.1	9.3
6.2	6.0	3.1	1.0	4.2
3.0	9.1	7.2	4.0	7.6

13.3 En el artículo "Shelf-Space Strategy in Retailing", que se publicó en *Proceedings: Southern Marketing Association*, se investigó el efecto que tenía la altura de los anaqueles en los supermercados sobre las ventas de alimento enlatado para perro. Durante un periodo de 8 días se llevó a cabo un experimento en un supermercado pequeño acerca de las ventas de una marca de alimento para perro conocida como Arf y se utilizaron tres niveles de altura de anaquel: a las rodillas, a la cintura y a los ojos. Cada día se cambió al azar tres veces la altura del anaquel en la que estaba dicho alimento. Las secciones restantes de la góndola que contenía la marca dada se llenaban con una mezcla de marcas de comida canina, las cuales resultaban tanto familiares como desconocidas para los consumidores de esa área geográfica específica. Se presentan las ventas diarias, expresadas en cientos de dólares, del alimento Arf para las tres alturas del anaquel. Con base en los datos, ¿existe una diferencia significativa en el promedio de ventas diarias de dicho alimento, con base en la altura del anaquel? Utilice un nivel de significancia de 0.01.

Altura de anaquel		
Nivel de las rodillas	**Nivel de la cintura**	**Nivel de los ojos**
77	88	85
82	94	85
86	93	87
78	90	81
81	91	80
86	94	79
77	90	87
81	87	93

13.4 La inmovilización de los venados silvestres de cola blanca usando tranquilizantes da a los investigadores la oportunidad de estudiarlos de cerca y obtener información fisiológica valiosa. En el estudio denominado *Influence of Physical Restraint and Restraint Facilitating Drugs on Blood Measurements of White-Tailed Deer and Other Selected Mammals*, realizado en Virginia Tech, los biólogos de la vida silvestre probaron el tiempo del "derribamiento" (el periodo transcurrido entre la inyección y la inmovilización) de tres sustancias inmovilizadoras distintas. En este caso la inmovilización se define como el punto en que el animal ya no tiene control muscular suficiente para permanecer de pie. Se asignaron 30 venados machos de cola blanca al azar a cada uno de tres tratamientos. El grupo *A* recibió 5 miligramos de cloruro de sucinilcolina líquida (SCC); al grupo *B* se le suministraron 8 miligramos de SCC en polvo; y al grupo *C*, 200 miligramos de hidrocloruro de fenciclidina. A continuación se presentan los tiempos de derribamiento, en minutos. Haga un análisis de varianza a un nivel de significancia de 0.01 y determine si el tiempo promedio de derribamiento es o no igual para las tres sustancias.

Grupo		
A	***B***	***C***
11	10	4
5	7	4
14	16	6
7	7	3
10	7	5
7	5	6
23	10	8
4	10	3
11	6	7
11	12	3

13.5 La enzima mitocondrial transhidrogenasa NADPH:NAD, de la tenia de la rata común (*Hymenolepiasis diminuta*) cataliza el hidrógeno en la transferencia de NADPH a NAD, lo que produce NADH. Se sabe que esta enzima desempeña un papel vital en el metabolismo anaerobio de la tenia, y recientemente se planteó la hipótesis de que podría servir como una bomba de intercambio de protones, es decir, para transferir protones a través de la membrana mitocondrial. Un estudio sobre el *Effect of Various Substrate Concentrations on the Conformational Variation of the NADPH:NAD Transhydrogenase of Hymenolepiasis diminuta* llevado a cabo por la Bowling Green State University, se diseñó para evaluar la capacidad de dicha enzima para sufrir cambios en su conformación o su forma. Podría considerarse que los cambios en la actividad específica de la enzima ocasionados por las variaciones en la concentración de NADP sustentan la teoría del cambio de conformación. La enzima en cuestión se localiza en la membrana interior de las mitocondrias de la tenia. Se homogeneizaron las tenias y se aisló la enzima mediante una serie de centrifugaciones. Después se agregaron diferentes concentraciones de NADP a la solución de enzima aislada y la mezcla se incubó durante tres minutos en un baño de agua a 56°C. Luego, se analizó la enzima con un espectrómetro de rayo dual y se calcularon los resultados que se presentan a continuación, en términos de la actividad específica de la enzima, en nanomoles por minuto por miligramo de proteína. Pruebe la hipótesis de que la actividad específica promedio es la misma para las cuatro concentraciones, a un nivel de significancia de 0.01.

Concentración de NADP (nm)				
0	**80**	**160**	**360**	
11.01	11.38	11.02	6.04	10.31
12.09	10.67	10.67	8.65	8.30
10.55	12.33	11.50	7.76	9.48
11.26	10.08	10.31	10.13	8.89
			9.36	

13.6 Un estudio midió la tasa de sorción (ya sea absorción o adsorción) de tres tipos diferentes de solventes químicos orgánicos. Estos solventes se utilizan para limpiar partes industriales metálicas, y son desechos potencialmente riesgosos. Se probaron muestras independientes de solventes de cada tipo y se registraron sus tasas de sorción como un porcentaje molar. (Véase McClave, Dietrich y Sincich, 1997).

Aromáticos		Cloroalcalinos		Ésteres		
1.06	0.95	1.58	1.12	0.29	0.43	0.06
0.79	0.65	1.45	0.91	0.06	0.51	0.09
0.82	1.15	0.57	0.83	0.44	0.10	0.17
0.89	1.12	1.16	0.43	0.55	0.53	0.17
1.05				0.61	0.34	0.60

¿Existe una diferencia significativa en la tasa promedio de sorción de los tres solventes? Utilice un valor *P* para sus conclusiones. ¿Qué solvente usaría?

13.7 Se ha demostrado que el fertilizante fosfato amoniacal de magnesio, $MgNH_4PO_4$, es un proveedor eficaz de los nutrientes necesarios para el crecimiento de las plantas. Los compuestos que suministra son muy solubles en agua, lo cual permite su aplicación directa sobre la superficie del suelo o que se mezcle con el sustrato de crecimiento durante el proceso de encapsu-

lamiento. Se efectuó un estudio denominado *Effect of Magnesium Ammonium Phosphate on Height of Chrysanthemums* en George Mason University para determinar el nivel óptimo posible de la fertilización con base en la mejoría de la respuesta de crecimiento vertical del crisantemo. Se dividieron 40 semillas de crisantemo en 4 grupos de diez plantas cada uno. Se sembró cada una en una maceta similar que contenía un medio uniforme de crecimiento. Se agregó a cada grupo de plantas una concentración cada vez mayor de $MgNH_4PO_4$, medido en gramos por bushel. Los cuatro grupos de plantas se cultivaron durante cuatro semanas en condiciones uniformes en un invernadero. A continuación se presentan los tratamientos y los cambios respectivos de sus alturas, medidas en centímetros:

Tratamiento							
50 g/bu		**100 g/bu**		**200 g/bu**		**400 g/bu**	
13.2	12.4	16.0	12.6	7.8	14.4	21.0	14.8
12.8	17.2	14.8	13.0	20.0	15.8	19.1	15.8
13.0	14.0	14.0	23.6	17.0	27.0	18.0	26.0
14.2	21.6	14.0	17.0	19.6	18.0	21.1	22.0
15.0	20.0	22.2	24.4	20.2	23.2	25.0	18.2

A un nivel de significancia de 0.05, ¿podría concluirse que concentraciones diferentes de $MgNH_4PO_4$ afectan la altura promedio que alcanzan los crisantemos? ¿Qué cantidad del fertilizante parece ser la mejor?

13.8 Para el conjunto de datos del ejercicio 13.7 use la prueba de Bartlett para probar si las varianzas son iguales. Utilice $\alpha = 0.05$.

13.9 Utilice la prueba de Bartlett a un nivel de significancia de 0.01 para probar la homogeneidad de las varianzas en el ejercicio 13.5 de la página 519.

13.10 Utilice la prueba de Cochran a un nivel de significancia de 0.01 para probar la homogeneidad de las varianzas en el ejercicio 13.4 de la página 519.

13.11 Utilice la prueba de Bartlett a un nivel de significancia de 0.05 para probar la homogeneidad de las varianzas en el ejercicio 13.6 de la página 519.

13.5 Comparaciones de un grado de libertad

El análisis de varianza en la clasificación de un solo factor, o experimento de un solo factor, como se le denomina con frecuencia, tan sólo indica si puede rechazarse o no la hipótesis de medias de tratamientos iguales. Por lo general, el experimentador preferiría efectuar un análisis más profundo. Como ilustración, en el ejemplo 13.1, mediante el rechazo de la hipótesis nula, concluimos que las medias no son iguales, pero aún no sabemos en dónde residen las diferencias entre los agregados. Es probable que el ingeniero intuya de antemano que los agregados 1 y 2 deberían poseer propiedades similares de absorción, al igual que los agregados 3 y 5. Sin embargo, sería interesante estudiar las diferencias entre los dos grupos. Así, parece apropiado probar las hipótesis

$$H_0: \ \mu_1 + \mu_2 - \mu_3 - \mu_5 = 0,$$
$$H_1: \ \mu_1 + \mu_2 - \mu_3 - \mu_5 \neq 0.$$

Se observa que la hipótesis es una función lineal de las medias de la población, en las cuales los coeficientes suman cero.

Definición 13.1: Cualquier función lineal de la forma

$$\omega = \sum_{i=1}^{k} c_i \mu_i,$$

donde $\sum_{i=1}^{k} c_i = 0$ se llama **comparación** o **contraste** en las medias de los tratamientos.

Con frecuencia el experimentador puede hacer comparaciones múltiples probando la significancia de los contrastes de las medias de los tratamientos, es decir, probando una hipótesis del siguiente tipo:

Hipótesis para un contraste

$$H_0: \sum_{i=1}^{k} c_i \mu_i = 0,$$

$$H_1: \sum_{i=1}^{k} c_i \mu_i \neq 0,$$

donde $\sum_{i=1}^{k} c_i = 0$.

La prueba se efectúa calculando primero un contraste similar de las medias de las muestras,

$$w = \sum_{i=1}^{k} c_i \bar{y}_{i.}\,.$$

Como $\overline{Y}_1$, $\overline{Y}_2$,..., $\overline{Y}_k$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \ldots, \mu_k$ y varianzas $\sigma_1^2/n_1, \sigma_2^2/n_2, \ldots, \sigma_k^2/n_k$, respectivamente, el teorema 7.11 nos garantiza que w es un valor de la variable aleatoria normal W con

$$\text{media } \mu_W = \sum_{i=1}^{k} c_i \mu_i \text{ y varianza } \sigma_W^2 = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i}.$$

Por lo tanto, cuando H_0 es verdadera, $\mu_W = 0$ y, según el ejemplo 7.5, el estadístico

$$\frac{W^2}{\sigma_W^2} = \frac{\left(\sum_{i=1}^{k} c_i \bar{Y}_{i.}\right)^2}{\sigma^2 \sum_{i=1}^{k} (c_i^2/n_i)}$$

se distribuye como una variable aleatoria chi cuadrada con 1 grado de libertad.

Estadístico de prueba para probar un contraste

Nuestra hipótesis se prueba a un nivel de significancia α calculando

$$f = \frac{\left(\sum_{i=1}^{k} c_i \bar{y}_{i.}\right)^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{\left[\sum_{i=1}^{k} (c_i Y_{i.}/n_i)\right]^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{SCw}{s^2}.$$

Aquí f es un valor de la variable aleatoria F que tiene distribución F con 1 y $N-k$ grados de libertad.

Cuando los tamaños de las muestras son iguales a n,

$$SCw = \frac{\left(\sum_{i=1}^{k} c_i Y_{i.}\right)^2}{n \sum_{i=1}^{k} c_i^2}.$$

La cantidad SCw, que se denomina **suma de cuadrados de los contrastes**, indica la parte de la SCT que se explica por el contraste en cuestión.

Esta suma de cuadrados se empleará para probar la hipótesis de que

$$\sum_{i=1}^{k} c_i \mu_i = 0.$$

Con frecuencia es de interés probar contrastes múltiples, en particular contrastes que son linealmente independientes u ortogonales. Como resultado, se vuelve necesaria la siguiente definición:

Definición 13.2: Se dice que los dos contrastes

$$\omega_1 = \sum_{i=1}^{k} b_i \mu_i \qquad \text{y} \qquad \omega_2 = \sum_{i=1}^{k} c_i \mu_i$$

son **ortogonales**, si $\sum_{i=1}^{k} b_i c_i / n_i = 0$, o bien, cuando las n_i son iguales a n, si

$$\sum_{i=1}^{k} b_i c_i = 0.$$

Si ω_1 y ω_2 son ortogonales, entonces las cantidades SCw_1 y SCw_2 son componentes de SCT, cada una con un solo grado de libertad. La suma de cuadrados de los tratamientos con $k - 1$ grados de libertad se puede dividir en, a lo sumo, $k - 1$ sumas de cuadrados de contrastes independientes con un solo grado de libertad que satisfacen la identidad

$$SCT = SCw_1 + SCw_2 + \cdots + SCw_{k-1},$$

si los contrastes son ortogonales entre sí.

Ejemplo 13.4: Remítase al ejemplo 13.1 y calcule la suma de cuadrados de los contrastes que corresponden a los contrastes ortogonales

$$\omega_1 = \mu_1 + \mu_2 - \mu_3 - \mu_5, \qquad \omega_2 = \mu_1 + \mu_2 + \mu_3 - 4\mu_4 + \mu_5,$$

y efectúe las pruebas de significancia adecuadas. En este caso es de interés *a priori* comparar los dos grupos (1, 2) y (3, 5). Un contraste importante e independiente consiste en realizar la comparación entre el conjunto de agregados (1, 2, 3, 5) y el agregado 4.

***Solución*:** Es evidente que los dos contrastes son ortogonales, puesto que

$$(1)(1) + (1)(1) + (-1)(1) + (0)(-4) + (-1)(1) = 0.$$

El segundo contraste indica una comparación entre los agregados (1, 2, 3 y 5) y el agregado 4. Podemos escribir dos contrastes adicionales ortogonales a los dos primeros, es decir:

$$\omega_3 = \mu_1 - \mu_2 \text{ (agregado 1 contra agregado 2)},$$
$$\omega_4 = \mu_3 - \mu_5 \text{ (agregado 3 contra agregado 5)}.$$

De los datos de la tabla 13.1, se tiene que

$$SCw_1 = \frac{(3320 + 3416 - 3663 - 3664)^2}{6[(1)^2 + (1)^2 + (-1)^2 + (-1)^2]} = 14,553,$$

$$SCw_2 = \frac{[3320 + 3416 + 3663 + 3664 - 4(2791)]^2}{6[(1)^2 + (1)^2 + (1)^2 + (1)^2 + (-4)^2]} = 70,035.$$

En la tabla 13.5 se presenta un análisis de varianza más extenso. Se observa que las dos sumas de cuadrados de los contrastes explican casi toda la suma de cuadrados de los agregados. Existe una diferencia significativa entre las propiedades de absorción de los agregados, y el contraste ω_1 es significativo marginalmente. Sin embargo, el valor f de 14.12 para ω_2 es muy significativo, y se rechaza la hipótesis

$$H_0: \ \mu_1 + \mu_2 + \mu_3 + \mu_5 = 4\mu_4$$

Tabla 13.5: Análisis de varianza usando contrastes ortogonales

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Agregados	85,356	4	21,339	4.30
(1, 2) *vs*. (3,5)	14,553	1	14,553	2.93
(1, 2, 3, 5) *vs*. 4	70,035	1	70,035	14.12
Error	124,021	25	4961	
Total	209,377	29		

Los contrastes ortogonales permiten al profesional dividir la variación del tratamiento en componentes independientes. Por lo general el experimentador tiene interés en hacer ciertos contrastes. Eso ocurrió en nuestro ejemplo, donde las consideraciones *a priori* sugerían que los agregados (1, 2) y (3, 5) constituían grupos distintos con propiedades diferentes de absorción, un planteamiento que no obtuvo mucho respaldo con la prueba de significancia. Sin embargo, la segunda comparación apoyó la conclusión de que el agregado 4 parecía "destacar" de los demás. En este caso no fue necesaria la partición completa de *SCT*, dado que dos de las cuatro comparaciones independientes posibles explicaban la mayor parte de la variación en los tratamientos.

En la figura 13.4 se presenta un procedimiento GLM del programa *SAS*, que proporciona un conjunto completo de contrastes ortogonales. Observe que la suma de cuadrados de los cuatro contrastes se agrega a la suma de cuadrados de los agregados. Asimismo, note que los últimos dos contrastes (1 contra 2, 3 contra 5) revelan comparaciones insignificantes.

13.6 Comparaciones múltiples

El análisis de varianza es un procedimiento poderoso para probar la homogeneidad de un conjunto de medias. No obstante, si se rechazara la hipótesis nula y se aceptara la alternativa que se planteó (que no todas las medias son iguales), aún no se sabría cuáles de las medias de la población son iguales y cuáles son diferentes.

```
                              The GLM Procedure
Dependent Variable: moisture
                                  Sum of
 Source                DF       Squares    Mean Square   F Value   Pr > F
 Model                  4    85356.4667     21339.1167      4.30   0.0088
 Error                 25   124020.3333      4960.8133
 Corrected Total       29   209376.8000

          R-Square       Coeff Var       Root MSE     moisture Mean
          0.407669        12.53703       70.43304           561.8000

 Source               DF      Type I SS      Mean Square      F Value     Pr > F
 aggregate             4    85356.46667      21339.11667         4.30     0.0088

 Source               DF    Type III SS      Mean Square      F Value     Pr > F
 aggregate             4    85356.46667      21339.11667         4.30     0.0088

 Contrast             DF   Contrast SS      Mean Square      F Value     Pr > F
 (1,2,3,5) vs. 4       1    70035.00833      70035.00833        14.12     0.0009
 (1,2) vs. (3,5)       1    14553.37500      14553.37500         2.93     0.0991
 1 vs. 2           1          768.00000        768.00000         0.15     0.6973
 3 vs. 5           1            0.08333          0.08333         0.00     0.9968
```

Figura 13.4: Un conjunto de contrastes ortogonales.

A menudo es de interés efectuar varias **comparaciones por pares** (quizá todas las que sean posibles) entre los tratamientos. En realidad, una comparación por pares se puede ver como un contraste simple, es decir, una prueba de

$$H_0\colon\ \mu_i - \mu_j = 0,$$
$$H_1\colon\ \mu_i - \mu_j \neq 0,$$

para toda $i \neq j$. Hacer todas las comparaciones posibles por pares entre las medias puede ser muy benéfico cuando no se conocen *a priori* contrastes complejos particulares. Por ejemplo, suponga que se desea probar las hipótesis siguientes, con los datos de los agregados de la tabla 13.1:

$$H_0\colon\ \mu_1 - \mu_5 = 0,$$
$$H_1\colon\ \mu_1 - \mu_5 \neq 0.$$

La prueba se desarrolla usando una F, una t, o el método de los intervalos de confianza. Si se usa la t, se tiene que

$$t = \frac{\bar{y}_{1.} - \bar{y}_{5.}}{s\sqrt{2/n}},$$

donde s es la raíz cuadrada del cuadrado medio del error y $n = 6$ es el tamaño de la muestra por tratamiento. En este caso,

$$t = \frac{553.33 - 610.67}{\sqrt{4961}\sqrt{1/3}} = -1.41.$$

El valor P para la prueba t con 25 grados de libertad es 0.17. Así que no hay evidencia suficiente para rechazar H_0.

Relación entre *T* y *F*

Anteriormente se expuso el uso de una prueba t agrupada, junto con los lineamientos que se estudiaron en el capítulo 10. El estimado agrupado se tomó del cuadrado medio del error con el fin de aprovechar los grados de libertad que están agrupados en las cinco muestras. Además, probamos un contraste. El lector debería observar que si el valor t se eleva al cuadrado, el resultado tiene exactamente la misma forma que el valor de f para una prueba del contraste, analizada en la sección anterior. En efecto,

$$f = \frac{(\bar{y}_{1.} - \bar{y}_{5.})^2}{s^2(1/6 + 1/6)} = \frac{(553.33 - 610.67)^2}{4961(1/3)} = 1.988,$$

que es, por supuesto, t^2.

Método del intervalo de confianza para una comparación por pares

Es fácil resolver el mismo problema de una comparación por pares (o un contraste) usando el método del intervalo de confianza. Es claro que, si se calcula un intervalo de confianza del $100(1 - \alpha)\%$ sobre $\mu_1 - \mu_5$, se tiene que

$$\bar{y}_{1.} - \bar{y}_{5.} \pm t_{\alpha/2} s \sqrt{\frac{2}{6}},$$

donde $t_{\alpha/2}$ es el punto superior de $100(1 - \alpha/2)\%$ de una distribución t con 25 grados de libertad (grados de libertad que provienen de s^2). Esta conexión inmediata entre las pruebas de hipótesis y los intervalos de confianza debería ser evidente a partir de los análisis de los capítulos 9 y 10. La prueba de un contraste simple $\mu_1 - \mu_5$ no implica más que observar si el intervalo de confianza anterior cubre o no al cero. Al sustituir los números se tiene que el intervalo de confianza de 95%:

$$(553.33 - 610.67) \pm 2.060 \sqrt{4961} \sqrt{\frac{1}{3}} = -57.34 \pm 83.77.$$

Por consiguiente, como el intervalo de confianza cubre al cero, el contraste no es significativo. En otras palabras, no hay diferencia significativa entre las medias de los agregados 1 y 5.

Tasa de error por experimento

Se presentan muchas dificultades cuando el analista intenta hacer muchas o todas las comparaciones por pares posibles. Para el caso de k medias habrá, desde luego, $r = k(k - 1)/2$ comparaciones por pares posibles. Si se suponen comparaciones independientes, la tasa de error por experimento o tasa de error por familia, es decir, la probabilidad de un falso rechazo de al menos una de las hipótesis, es dada por $1 - (1 - \alpha)^r$, donde α es la probabilidad seleccionada del error tipo I para una comparación específica. Es claro que esta medida del error tipo I por experimento sería bastante grande. Por

ejemplo, aun si sólo hubiera 6 comparaciones, digamos, en el caso de 4 medias, y $\alpha = 0.05$, la tasa de experimento-juicio sería

$$1 - (0.95)^6 \approx 0.26.$$

Cuando se prueban muchas comparaciones por pares, por lo general existe la necesidad de hacer el contraste efectivo sobre una sola comparación más conservadora. Es decir, usando el método del intervalo de confianza, los intervalos de confianza serían mucho más anchos que $\pm t_{\alpha/2} s\sqrt{2/n}$ que se emplea para el caso de una sola comparación.

Prueba de Tukey

Hay varios métodos estándar para realizar comparaciones por pares que den credibilidad a la tasa del error tipo I. Aquí se analizarán e ilustrarán dos de ellos. El primero, denominado **procedimiento de Tukey**, permite la formación de intervalos de confianza del $100(1 - \alpha)\%$ simultáneos para todas las comparaciones por pares. El método se basa en la distribución del rango *estudentizado*. El punto apropiado del percentil es una función de α, k y v = grados de libertad para s^2. En la tabla A.12 se presenta una lista de puntos porcentuales superiores adecuados para $\alpha = 0.05$. El método de Tukey de comparaciones por pares implica encontrar una diferencia significativa entre las medias i y j ($i \neq j$) si $|\bar{y}_{i.} - \bar{y}_{j.}|$ excede a $q(\alpha, k, v)\sqrt{\frac{s^2}{n}}$.

El procedimiento de Tukey se ilustra con facilidad. Considere un ejemplo hipotético en el que se tienen 6 tratamientos en un diseño completamente aleatorizado de un solo factor, en el que se hacen 5 observaciones por tratamiento. Suponga que el cuadrado medio del error tomado de la tabla del análisis de varianza es $s^2 = 2.45$ (24 grados de libertad). Las medias muestrales están en orden ascendente,

$\bar{y}_{2.}$	$\bar{y}_{5.}$	$\bar{y}_{1.}$	$\bar{y}_{3.}$	$\bar{y}_{6.}$	$\bar{y}_{4.}$
14.50	16.75	19.84	21.12	22.90	23.20.

Con $\alpha = 0.05$, el valor de $q(0.05, 6, 24)$ es 4.37. Así, todas las diferencias absolutas se comparan con

$$4.37\sqrt{\frac{2.45}{5}} = 3.059.$$

Como resultado, las siguientes representan medias que, usando el procedimiento de Tukey, se encuentra que son significativamente diferentes:

4 y 1,	4 y 5,	4 y 2,	6 y 1,	6 y 5,
6 y 2,	3 y 5,	3 y 2,	1 y 5,	1 y 2.

¿De dónde proviene el nivel α en la prueba de Tukey?

Se mencionó brevemente el concepto de **intervalos de confianza simultáneos** que se emplean para el procedimiento de Tukey. El lector obtendrá una perspectiva útil del concepto de comparaciones múltiples, si comprende el significado de los intervalos de confianza simultáneos.

En el capítulo 9 vimos que, si se calcula un intervalo de confianza de 95% para, digamos, una media μ, entonces la probabilidad de que el intervalo cubra la media verdadera μ es 0.95.

Sin embargo, como vimos antes, para el caso de comparaciones múltiples la probabilidad efectiva de interés está ligada con la tasa de error por experimento, y debe hacerse énfasis en que los intervalos de confianza del tipo $\bar{y}_{i.} - \bar{y}_{j.} \pm q(\alpha, k, v)s\sqrt{1/n}$ no son independientes, ya que todos implican a s y muchos utilizan los mismos promedios, las $\bar{y}_{i.}$. A pesar de tales dificultades, si se utiliza la $q(0.05, k, v)$, el nivel de confianza simultáneo está controlado en un 95%. Lo mismo es cierto para $q(0.01, k, v)$, es decir, el nivel de confianza está controlado en un 99%. En el caso de $\alpha = 0.05$, hay una probabilidad de 0.05 de que se encuentre falsamente que al menos un par de mediciones son diferentes (falso rechazo de al menos una hipótesis nula). En el caso de $\alpha = 0.01$, la probabilidad correspondiente será 0.01.

Prueba de Duncan

El segundo procedimiento que se estudiará se denomina **procedimiento de Duncan** o **prueba de Duncan de rango múltiple**. Este procedimiento también se basa en el concepto general del rango estudentizado. El rango de cualquier subconjunto de p medias muestrales debe exceder cierto valor antes de que se encuentre que cualquiera de las p medias es diferente. Este valor recibe el nombre de **rango de menor significancia** para las p medias, y se denota por R_p, donde

$$R_p = r_p\sqrt{\frac{s^2}{n}}.$$

Los valores de la cantidad r_p, llamados **rango estudentizado de menor significancia**, dependen del nivel de significancia deseado y del número de grados de libertad del cuadrado medio del error. Estos valores se obtienen de la tabla A.13 para $p = 2, 3,..., 10$ medias.

Para ilustrar el procedimiento de prueba de rango múltiple, consideremos el ejemplo hipotético en el cual se comparan 6 tratamientos con 5 observaciones por tratamiento. Se trata del mismo ejemplo que se empleó para ilustrar la prueba de Tukey. Se obtiene R_p multiplicando cada r_p por 0.70. Los resultados de estos cálculos se resumen como sigue:

p	2	3	4	5	6
r_p	2.919	3.066	3.160	3.226	3.276
R_p	2.043	2.146	2.212	2.258	2.293

Si se comparan estos rangos de menor significancia con las diferencias en medias ordenadas, se llega a las conclusiones siguientes:

1. Como $\bar{y}_{4.} - \bar{y}_{2.} = 8.70 > R_6 = 2.293$, se concluye que μ_4 y μ_2 son significativamente distintas.
2. Si se comparan $\bar{y}_{4.} - \bar{y}_{5.}$ y $\bar{y}_{6.} - \bar{y}_{2.}$ con R_5, se concluye que μ_4 es significativamente mayor que μ_5 y que μ_6 es significativamente mayor que μ_2.
3. Si se comparan $\bar{y}_{4.} - \bar{y}_{1.}$, $\bar{y}_{6.} - \bar{y}_{5.}$ y $\bar{y}_{3.} - \bar{y}_{2.}$ con R_4, se concluye que cada diferencia es significativa.
4. Si se comparan $\bar{y}_{4.} - \bar{y}_{3.}$, $\bar{y}_{6.} - \bar{y}_{1.}$, $\bar{y}_{3.} - \bar{y}_{5.}$ y $\bar{y}_{1.} - \bar{y}_{2.}$ con R_3, se encuentra que todas las diferencias son significativas excepto para $\mu_4 - \mu_3$. Por lo tanto, μ_3, μ_4 y μ_6 constituyen un subconjunto de medias homogéneas.
5. Si se comparan $\bar{y}_{3.} - \bar{y}_{1.}$, $\bar{y}_{1.} - \bar{y}_{5.}$ y $\bar{y}_{5.} - \bar{y}_{2.}$ con R_2, se concluye que sólo μ_3 y μ_1 no son significativamente distintas.

Se acostumbra resumir las conclusiones anteriores con el dibujo de una línea debajo de cualquier subconjunto de medias adyacentes que no sean significativamente diferentes. Así, tenemos

$\bar{y}_{2.}$	$\bar{y}_{5.}$	$\bar{y}_{1.}$	$\bar{y}_{3.}$	$\bar{y}_{6.}$	$\bar{y}_{4.}$
14.50	16.75	19.84	21.12	22.90	23.20

Es evidente que en este caso los resultados con los procedimientos de Tukey y Duncan son muy similares. El procedimiento de Tukey no detectó ninguna diferencia entre 2 y 5; mientras que el de Duncan sí lo hizo.

Prueba de Dunnett: comparación de tratamientos con un control

En muchos problemas científicos y de ingeniería no nos interesa hacer inferencias acerca de todas las comparaciones posibles entre las medias de los tratamientos del tipo $\mu_i - \mu_j$. En vez de ello es frecuente que el experimento dicte la necesidad de comparar simultáneamente cada *tratamiento* con un *control*. Un procedimiento de prueba desarrollado por C. W. Dunnett determina diferencias significativas entre cada media de tratamiento y el control, con un solo nivel conjunto de significancia α. Para ilustrar el procedimiento de Dunnett, se considerarán los datos experimentales de la tabla 13.6 para la clasificación de un solo factor, donde se estudió el efecto de tres catalizadores sobre el producto de una reacción. Como control se emplea un cuarto tratamiento en el que no se aplica un catalizador.

Tabla 13.6: Producto de una reacción

Control	Catalizador 1	Catalizador 2	Catalizador 3
50.7	54.1	52.7	51.2
51.5	53.8	53.9	50.8
49.2	53.1	57.0	49.7
53.1	52.5	54.1	48.0
52.7	54.0	52.5	47.2
$\bar{y}_{0.} = 51.44$	$\bar{y}_{1.} = 53.50$	$\bar{y}_{2.} = 54.04$	$\bar{y}_{3.} = 49.38$

En general, se desea probar las k hipótesis

$$\left.\begin{array}{l} H_0\colon\ \mu_0 = \mu_i \\ H_1\colon\ \mu_0 \neq \mu_i \end{array}\right\} \qquad i = 1, 2, \ldots, k,$$

donde μ_0 representa el producto medio para la población de medidas en que se utiliza el control. Como se mencionó en la sección 13.3, se espera que las suposiciones habituales del análisis de varianza sigan siendo válidas. Para probar la hipótesis nula especificada con H_0 en comparación con alternativas bilaterales para una situación experimental donde existen k tratamientos, sin incluir el control, y n observaciones por tratamiento, primero calculamos los valores

$$d_i = \frac{\bar{y}_{i.} - \bar{y}_{0.}}{\sqrt{2s^2/n}}, \qquad i = 1, 2, \ldots, k.$$

Como antes, la varianza muestral s^2 se obtiene a partir del cuadrado medio del error en el análisis de varianza. Ahora bien, la región crítica para rechazar H_0 a un nivel de significancia α, se establece con la desigualdad

$$|d_i| > d_{\alpha/2}(k, v),$$

donde v es el número de grados de libertad para el cuadrado medio del error. Los valores de la cantidad $d_{\alpha/2}(k, v)$ para una prueba de dos colas se incluyen en la tabla A.14 para $\alpha = 0.05$ y $\alpha = 0.01$, para diversos valores de k y v.

Ejemplo 13.5: Para los datos de la tabla 13.6, pruebe la hipótesis que compara cada catalizador con el control, usando alternativas bilaterales. Como nivel de significancia conjunto elija $\alpha = 0.05$.

***Solución*:** El cuadrado medio del error con 16 grados de libertad se obtiene de la tabla de análisis de varianza, usando todos los tratamientos $k + 1$. El cuadrado medio del error es dado por

$$s^2 = \frac{36.812}{16} = 2.30075 \text{ y } \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{(2)(2.30075)}{5}} = 0.9593.$$

Entonces,

$$d_1 = \frac{53.50 - 51.44}{0.9593} = 2.147, \quad d_2 = \frac{54.04 - 51.44}{0.9593} = 2.710,$$

$$d_3 = \frac{49.38 - 51.44}{0.9593} = -2.147.$$

De la tabla A.14 el valor crítico para $\alpha = 0.05$ resulta ser $d_{0.025}(3, 16) = 2.59$. Como $|d_1| < 2.59$ y $|d_3| < 2.59$, se concluye que tan sólo la producción media para el catalizador 2 es significativamente diferente de la respuesta media de la reacción utilizando el control.

Muchas aplicaciones prácticas imponen la necesidad de una prueba de una cola para comparar los tratamientos con un control. En efecto, cuando un farmacólogo está interesado en el efecto de varias dosis de un medicamento sobre el nivel del colesterol, y su control consiste en una dosis de cero, sería interesante determinar si cada dosis produce una reducción significativamente mayor que la del control. En la tabla A.15 se presentan los valores críticos de $d_\alpha(k, v)$ para alternativas unilaterales.

Ejercicios

13.12 Considere los datos del ejercicio de repaso 13.45 de la página 555. Efectúe pruebas de significancia sobre los siguientes contrastes:

a) *B* contra *A*, *C* y *D*;
b) *C* contra *A* y *D*;
c) *A* contra *D*.

13.13 El propósito del estudio *The Incorporation of a Chelating Agent into a Flame Retardant Finish of a Cotton Flannelette and the Evaluation of Selected Fabric Properties*, llevado a cabo en Virginia Tech, fue evaluar el uso de un agente quelante como parte del acabado retardante del fuego de la franela de algodón, determinando sus efectos en la inflamabilidad después de lavar la tela en condiciones específicas. Se prepararon dos baños, uno con celulosa de carboximetilo y otro sin ella. Se lavaron 12 piezas de tela 5 veces en el baño I, y otras 12 piezas se lavaron 10 veces en el baño I. Esto se repitió con 24 piezas adicionales de tela en el baño II. Después de los lavados se midieron las longitudes quemadas de la tela, así como los tiempos de combustión. Por conveniencia, definamos los siguientes tratamientos:

Tratamiento 1: 5 lavados en el baño I,

Tratamiento 2: 5 lavados en el baño II,

Tratamiento 3: 10 lavados en el baño I,

Tratamiento 4: 10 lavados en el baño II.

Los registros del tiempo de combustión, en segundos, son los siguientes:

Tratamiento			
1	**2**	**3**	**4**
13.7	6.2	27.2	18.2
23.0	5.4	16.8	8.8
15.7	5.0	12.9	14.5
25.5	4.4	14.9	14.7
15.8	5.0	17.1	17.1
14.8	3.3	13.0	13.9
14.0	16.0	10.8	10.6
29.4	2.5	13.5	5.8
9.7	1.6	25.5	7.3
14.0	3.9	14.2	17.7
12.3	2.5	27.4	18.3
12.3	7.1	11.5	9.9

a) Efectúe un análisis de varianza a un nivel de significancia de 0.01, y determine si hay diferencias significativas entre las medias de los tratamientos.

b) Use contrastes de un solo grado de libertad con $\alpha = 0.01$ para comparar el tiempo medio de combustión del tratamiento 1 en comparación con el tratamiento 2, y también del tratamiento 3 en comparación con el 4.

13.14 El Departamento de Alimentación y Nutrición Humana de Virginia Tech realizó el estudio *Loss of Nitrogen Through Sweat by Preadolescent Boys Consuming Three Levels of Dietary Protein* para determinar la pérdida de nitrógeno por transpiración con varios niveles dietéticos de proteínas. En el experimento participaron 12 hombres preadolescentes cuyas edades iban de 7 años 8 meses a 9 años 8 meses, y a quienes de les calificó de clínicamente saludables. Cada muchacho estuvo sujeto a una de tres dietas controladas en las cuales consumía 29, 54 u 84 gramos de proteínas por día. Los siguientes datos representan la pérdida de nitrógeno corporal por transpiración, en miligramos, recabados durante los dos días últimos del periodo de experimentación:

Nivel de proteínas		
29 gramos	**54 gramos**	**84 gramos**
190	318	390
266	295	321
270	271	396
	438	399
	402	

a) Realice un análisis de varianza a un nivel de significancia de 0.05, para demostrar que las pérdidas medias de nitrógeno por transpiración son diferentes con los tres niveles de proteínas.

b) Utilice una prueba de Tukey para determinar cuáles niveles de proteínas difieren significativamente entre sí en la pérdida media de nitrógeno.

13.15 Utilice la prueba de Tukey a un nivel de significancia de 0.05, para analizar las medias de las 5 marcas diferentes de tabletas para el dolor de cabeza del ejercicio 13.2 de la página 518.

13.16 Se realizó una investigación para determinar la fuente de reducción en el rendimiento de cierto producto químico. Se sabía que la pérdida en el rendimiento ocurría en el licor madre, es decir, el material eliminado en la etapa de filtración. Se pensaba que mezclas distintas del material original podrían ocasionar reducciones diferentes del rendimiento en la etapa de licor madre. A continuación se presentan los resultados de la reducción porcentual para tres lotes de cada una de cuatro mezclas seleccionadas con anterioridad.

Mezcla			
1	**2**	**3**	**4**
25.6	25.2	20.8	31.6
24.3	28.6	26.7	29.8
27.9	24.7	22.2	34.3

a) Haga el análisis de varianza al nivel de significancia $\alpha = 0.05$.

b) Utilice la prueba de Duncan de rango múltiple para determinar cuáles mezclas difieren.

c) Resuelva el inciso *b* usando la prueba de Tukey.

13.17 En el estudio, denominado *An Evaluation of the Removal Method for Estimating Benthic Populations and Diversity*, realizado por Virginia Tech en el río Jackson, se emplearon 5 procedimientos distintos de muestreo para determinar los conteos de especies. Se seleccionaron 20 muestras al azar y los 5 procedimientos de muestreo se repitieron 4 veces. Se registraron los siguientes conteos de especies:

Procedimiento de muestreo				
Disminución	**De Hess modificado**	**Remoción del sustrato de Surber**	**Kicknet**	**Kicknet**
85	75	31	43	17
55	45	20	21	10
40	35	9	15	8
77	67	37	27	15

a) ¿Hay alguna diferencia significativa en el conteo promedio de especies para los distintos procedimientos de muestreo? Use un valor *P* en su conclusión.

b) Emplee una prueba de Tukey con $\alpha = 0.05$ para determinar cuáles procedimientos de muestreo difieren.

13.18 Los siguientes datos son valores de presión (psi) en un resorte de torsión para valores distintos del ángulo entre las vueltas del resorte en posición libre.

Ángulo					
67	**71**	**75**		**79**	**83**
83	84	86	87	89	90
85	85	87	87	90	92
	85	88	88	90	
	86	88	88	91	
	86	88	89		
	87	90			

Calcule un análisis de varianza de un solo factor para este experimento y plantee sus conclusiones acerca del efecto que tiene el ángulo sobre la presión en el resorte. (Tomado de C. R. Hicks, *Fundamental Concepts in the Design of Experiments*, Holt, Rinehart y Winston, Nueva York, 1973).

13.19 Se sospecha que la temperatura del ambiente en que se activan las baterías afecta su vida. Se probaron 30 baterías homogéneas, seis a cada una de cinco temperaturas, y los datos se presentan a continuación (vida activada en segundos). Analice e interprete los datos. (Tomado de C. R. Hicks, *Fundamental Concepts in Design of Experiments*, Holt, Rinehart y Winston, Nueva York, 1973.)

Temperatura(°C)				
0	**25**	**50**	**75**	**100**
55	60	70	72	65
55	61	72	72	66
57	60	72	72	60
54	60	68	70	64
54	60	77	68	65
56	60	77	69	65

13.20 La tabla siguiente (tomada de A. Hald, *Statistical Theory with Engineering Applications*, John Wiley & Sons, Nueva York, 1952) proporciona las resistencias a la tensión (en desviaciones desde 340) para conductores extraídos de nueve cables que deben usarse para una red de alto voltaje. Cada cable está constituido por 12 conductores. Se desea saber si las resistencias medias de los conductores en los nueve cables son las mismas. Si los cables son diferentes, ¿cuáles son los que difieren? Utilice un valor P en su análisis de varianza.

Cable	Resistencia a la tensión											
1	5	−13	−5	−2	−10	−6	−5	0	−3	2	−7	−5
2	−11	−13	−8	8	−3	−12	−12	−10	5	−6	−12	−10
3	0	−10	−15	−12	−2	−8	−5	0	−4	−1	−5	−11
4	−12	4	2	10	−5	−8	−12	0	−5	−3	−3	0
5	7	1	5	0	10	6	5	2	0	−1	−10	−2
6	1	0	−5	−4	−1	0	2	5	1	−2	6	7
7	−1	0	2	1	−4	2	7	5	1	0	−4	2
8	−1	0	7	5	10	8	1	2	−3	6	0	5
9	2	6	7	8	15	11	−7	7	10	7	8	1

13.21 La salida de resultados que se presenta en la figura 13.5 de la página 532 proporciona información sobre la prueba de Duncan para los datos de los agregados del ejemplo 13.1 obtenidos con la función PROC GLM del programa *SAS*. Saque conclusiones sobre las comparaciones por pares usando los resultados de la prueba de Duncan.

13.22 Realice la prueba de Duncan para comparaciones por pares con los datos del ejercicio 13.6 de la página 519. Comente los resultados.

13.23 En un experimento biológico se emplearon 4 concentraciones de cierto producto químico para mejorar el crecimiento de cierto tipo de planta con el paso del tiempo. Se utilizaron cinco plantas con cada concentración y se midió su crecimiento, en centímetros. Se obtuvieron los siguientes datos y también se aplicó un control (ausencia de producto químico)

	Concentración			
Control	**1**	**2**	**3**	**4**
6.8	8.2	7.7	6.9	5.9
7.3	8.7	8.4	5.8	6.1
6.3	9.4	8.6	7.2	6.9
6.9	9.2	8.1	6.8	5.7
7.1	8.6	8.0	7.4	6.1

Utilice una prueba bilateral de Dunnett a un nivel de significancia de 0.05 para comparar de manera simultánea las concentraciones con el control.

13.24 La estructura financiera de una empresa consiste en la forma en que sus activos se dividen en capital y deuda, y el apalancamiento financiero se refiere al porcentaje de activos financiados con endeudamiento. En el artículo *The Effect of Financial Leverage on Return*, Tai Ma, de Virginia Tech, afirma que es posible utilizar el apalancamiento financiero para incrementar la tasa de rendimiento sobre el capital. Dicho de otra manera, los accionistas pueden recibir rendimientos más elevados sobre el capital propio con la misma cantidad de inversión si usan apalancamiento financiero. Los siguientes datos muestran las tasas de rendimiento sobre el capital utilizando 3 niveles distintos de apalancamiento financiero, así como un nivel de control (deuda igual a cero) para 24 empresas seleccionadas al azar.

Apalancamiento financiero			
Control	**Bajo**	**Medio**	**Alto**
2.1	6.2	9.6	10.3
5.6	4.0	8.0	6.9
3.0	8.4	5.5	7.8
7.8	2.8	12.6	5.8
5.2	4.2	7.0	7.2
2.6	5.0	7.8	12.0

Fuente: Standard & Poor's *Machinery Industry Survey*, 1975.

a) Haga el análisis de varianza a un nivel de significancia de 0.05.

b) Use una prueba de Dunnett a un nivel de significancia de 0.01, para determinar si las tasas medias de rendimiento sobre el capital son más elevadas con los niveles bajo, medio y alto de apalancamiento financiero que con el nivel de control.

```
                    The GLM Procedure
          Duncan's  Multiple  Range  Test  for moisture
NOTE: This test controls the Type I comparisonwise error rate,
                not the experimentwise error rate.
              Alpha                          0.05
              Error  Degrees  of  Freedom      25
              Error  Mean  Square          4960.813
  Number  of  Means          2          3          4          5
  Critical  Range        83.75      87.97      90.69      92.61

Means  with  the  same  letter  are  not  significantly  different.
        Duncan  Grouping          Mean       N     aggregate
                        A         610.67       6     5
                        A
                        A         610.50       6     3
                        A
                        A         569.33       6     2
                        A
                        A         553.33       6     1

                        B         465.17       6     4
```

Figura 13.5: Salida de resultados del *SAS* para el ejercicio 13.21.

13.7 Comparación de un conjunto de tratamientos en bloques

En la sección 13.2 estudiamos la idea de la formación de bloques, es decir, de aislar conjuntos de unidades experimentales que son razonablemente homogéneas y asignarles tratamientos de forma aleatoria. Ésta es una extensión del concepto de "formar pares" que se analizó en los capítulos 9 y 10, y se hace para reducir el error experimental, ya que las unidades en un bloque tienen más características comunes que las unidades localizadas en diferentes bloques.

El lector no debería considerar los bloques como un segundo factor, aunque ésa sea una forma tentadora de visualizar el diseño. De hecho, el factor principal (los tratamientos) aún lleva el peso mayor del experimento. Las unidades experimentales siguen siendo la fuente del error, igual que en el diseño completamente aleatorizado. Con la formación de bloques simplemente tratamos a dichas unidades de manera más sistemática. De ese modo, se dice que la aleatoriedad tiene restricciones. Antes de iniciar el estudio de la formación de bloques revisaremos dos ejemplos de un **diseño completamente aleatorizado**. El primer ejemplo es un experimento químico diseñado para determinar si hay una diferencia en la reacción media producida por cuatro catalizadores. Las muestras de los materiales que tienen que probarse se extraen de los mismos lotes de materias primas, a la vez que se mantienen constantes otras condiciones como la temperatura y concentración de los reactivos. En este caso, la hora del día en que se efectúan las corridas experimentales podría representar las unidades experimentales, y si el experimentador considera que es posible que haya un ligero efecto del tiempo, aleatorizaría la asignación

de los catalizadores a las corridas para contrarrestar la posible tendencia. Como un segundo ejemplo de dicho diseño, considere un experimento para comparar cuatro métodos para medir una propiedad física en particular de un fluido. Suponga que el proceso de muestreo es destructivo, es decir, que una vez que se ha medido una muestra de la sustancia usando un método, ya no puede medirse nuevamente con ningún otro. Si se decide hacer cinco mediciones con cada método, entonces se seleccionan al azar 20 muestras del material de un lote grande y se utilizan en el experimento para comparar los cuatro métodos de medición. Las unidades experimentales son las muestras seleccionadas al azar. Cualquier variación de una muestra a otra aparecerá en la variación del error, según se mida con s^2 en el análisis.

¿Cuál es el propósito de formar bloques?

Si la variación debida a la heterogeneidad en las unidades experimentales es tan grande que la sensibilidad para detectar diferencias de tratamiento se reduce debido a un valor aumentado de s^2, un plan mejor sería "bloquear" la variación debida a esas unidades y, por consiguiente, reducir la variación ajena a la que es explicada por bloques más pequeños o más homogéneos. Por ejemplo, suponga que en el ejemplo anterior de los catalizadores se supiera *a priori* que existe en definitiva un efecto significativo diario sobre el producto, y que es posible medir el producto para cuatro catalizadores en un día específico. En lugar de asignar los 4 catalizadores a las 20 corridas de prueba completamente al azar, se eligen, por ejemplo, 5 días y se prueba cada uno de los cuatro catalizadores cada día, asignándolos al azar a las corridas dentro de los días. De esta manera se elimina la variación diaria del análisis y, en consecuencia, el error experimental, que aún incluye cualquier tendencia temporal *dentro de los días*, representa con más precisión la variación aleatoria. A cada día se le denomina **bloque**.

El más directo de los diseños aleatorizados de bloques es aquel en el cual se asigna al azar un tratamiento a la vez a cada bloque. A un plan experimental como éste se le denomina **diseño de bloques completos aleatorizados (BCA)** y cada bloque constituye una sola réplica de los tratamientos.

13.8 Diseños de bloques completos aleatorizados

Un plan clásico para el diseño de bloques completos aleatorizados (BCA) usando tres mediciones en cuatro bloques es el siguiente:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
t_2	t_1	t_3	t_2
t_1	t_3	t_2	t_1
t_3	t_2	t_1	t_3

Las t denotan la asignación de cada uno de 3 tratamientos a los bloques. Desde luego, la asignación verdadera de los tratamientos a las unidades dentro de los bloques se hace al azar. Una vez que ha finalizado el experimento, los datos se pueden registrar como en el siguiente arreglo de 3×4:

Tratamiento	Bloque: 1	2	3	4
1	y_{11}	y_{12}	y_{13}	y_{14}
2	y_{21}	y_{22}	y_{23}	y_{24}
3	y_{31}	y_{32}	y_{33}	y_{34}

donde y_{11} representa la respuesta que se obtiene al utilizar el tratamiento 1 en el bloque 1, y_{12} es la respuesta que se obtiene al utilizar el tratamiento 1 en el bloque 2,..., y y_{34} es la respuesta que se obtiene al utilizar el tratamiento 3 en el bloque 4.

Ahora vamos a generalizar y a considerar el caso de k tratamientos asignados a b bloques. Los datos se pueden resumir tal como se observa en el arreglo rectangular de $k \times b$ de la tabla 13.7. Se supondrá que las y_{ij}, $i = 1, 2,..., k$ y $j = 1,2,..., b$, son valores de variables aleatorias independientes que tienen distribuciones normales con media μ_{ij} y varianza común σ^2.

Tabla 13.7: Arreglo de $k \times b$ para el diseño de BCA

	Bloque							
Tratamiento	**1**	**2**	…	***j***	…	***b***	**Total**	**Media**
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1b}	$T_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2b}	$T_{2.}$	$\bar{y}_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ib}	$T_{i.}$	$\bar{y}_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
k	y_{k1}	y_{k2}	…	y_{kj}	…	y_{kb}	$T_{k.}$	$\bar{y}_{k.}$
Total	$T_{.1}$	$T_{.2}$	…	$T_{.j}$	…	$T_{.b}$	$T_{..}$	
Media	$\bar{y}_{.1}$	$\bar{y}_{.2}$	…	$\bar{y}_{.j}$	…	$\bar{y}_{.b}$		$\bar{y}_{..}$

Sea $\mu_{i.}$ el promedio (en lugar del total) de las b medias de la población para el i-ésimo tratamiento. Es decir,

$$\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}, \text{ para } i = 1, \ldots, k.$$

De manera similar, el promedio de las medias de la población para el j-ésimo bloque, $\mu_{.j}$, es definido por

$$\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}, \text{ para } j = 1, \ldots, b$$

y el promedio de las bk medias de la población, μ, es definido por

$$\mu = \frac{1}{bk}\sum_{i=1}^{k}\sum_{j=1}^{b}\mu_{ij}.$$

Para determinar si parte de la variación de nuestras observaciones se debe a diferencias entre los tratamientos, se considera la siguiente prueba:

Hipótesis de medias iguales de los tratamientos

$$H_0: \ \mu_{1.} = \mu_{2.} = \cdots \mu_{k.} = \mu,$$
$$H_1: \text{ No todas las } \mu_{i.} \text{ son iguales.}$$

Modelo para el diseño BCA

Cada observación se puede escribir en la forma siguiente:

$$y_{ij} = \mu_{ij} + \epsilon_{ij},$$

donde ϵ_{ij} mide la desviación del valor observado y_{ij} de la media de la población μ_{ij}. La forma preferida de esta ecuación se obtiene sustituyendo

$$\mu_{ij} = \mu + \alpha_i + \beta_j,$$

donde α_i es, como antes, el efecto del *i*-ésimo tratamiento, y β_j es el efecto del *j*-ésimo bloque. Se supone que el tratamiento y los efectos de los bloques son aditivos. Por lo tanto, se puede escribir

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}.$$

Observe que el modelo se parece al de clasificación de un solo factor; la diferencia esencial es la introducción del efecto de bloque β_j. El concepto básico se parece mucho al de la clasificación de un solo factor, excepto que en el análisis debe tomarse en cuenta el efecto adicional debido a los bloques, ya que ahora la variación se controla de manera sistemática *en dos direcciones*. Si ahora imponemos las restricciones de que

$$\sum_{i=1}^{k} \alpha_i = 0 \qquad \text{y} \qquad \sum_{j=1}^{b} \beta_j = 0,$$

entonces

$$\mu_{i.} = \frac{1}{b} \sum_{j=1}^{b} (\mu + \alpha_i + \beta_j) = \mu + \alpha_i, \text{ para } i = 1, \ldots, k,$$

y

$$\mu_{.j} = \frac{1}{k} \sum_{i=1}^{k} (\mu + \alpha_i + \beta_j) = \mu + \beta_j, \text{ para } j = 1, \ldots, b.$$

La hipótesis nula de que las *k* medias de los tratamientos μ_i son iguales y, por lo tanto, iguales a μ, ahora es **equivalente a probar la hipótesis**:

$$H_0: \ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0,$$

$$H_1: \text{ Al menos una de las } \alpha_i \text{ no es igual a cero.}$$

Cada una de las pruebas de tratamientos se basará en una comparación de los estimados independientes de la varianza común poblacional σ^2. Esos estimados se

obtendrán separando la suma total de cuadrados de los datos en tres componentes mediante la siguiente identidad:

Teorema 13.3: Identidad de la suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

La demostración se deja como ejercicio para el lector.

La identidad de la suma de cuadrados podría presentarse simbólicamente mediante la ecuación

$$STC = SCT + SCB + SCE,$$

donde

$$STC = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = \text{suma total de cuadrados,}$$

$$SCT = b\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{suma de los cuadrados de los tratamientos,}$$

$$SCB = k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 = \text{suma de los cuadrados de los bloques,}$$

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 = \text{suma de los cuadrados del error.}$$

Si se sigue el procedimiento descrito en el teorema 13.2, donde se interpretó a las sumas de cuadrados como funciones de las variables aleatorias independientes, Y_{11}, Y_{12},..., Y_{kb}, se puede demostrar que los valores esperados de las sumas de los cuadrados de los tratamientos, los bloques y los errores son dados por

$$E(SCT) = (k-1)\sigma^2 + b\sum_{i=1}^{k}\alpha_i^2, \quad E(SCB) = (b-1)\sigma^2 + k\sum_{j=1}^{b}\beta_j^2,$$

$$E(SCE) = (b-1)(k-1)\sigma^2.$$

Como en el caso del problema de un solo factor, tenemos que el cuadrado medio del tratamiento es

$$s_1^2 = \frac{SCT}{k-1}.$$

Si los efectos del tratamiento $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, entonces s_1^2 es un estimado insesgado de σ^2. Sin embargo, si los efectos de los tratamientos no son todos iguales a cero, se tiene que:

Media cuadrada esperada del tratamiento

$$E\left(\frac{SCT}{k-1}\right) = \sigma^2 + \frac{b}{k-1}\sum_{i=1}^{k}\alpha_i^2$$

En este caso s_1^2 sobrestima σ^2. Un segundo estimado de σ^2, basado en $b - 1$ grados de libertad, es

$$s_2^2 = \frac{SCB}{b-1}.$$

El estimado s_2^2 es un estimado no sesgado de σ^2 si los efectos de los bloques $\beta_1 = \beta_2 = \cdots = \beta_b = 0$. Si los efectos de los bloques no son iguales a cero, entonces,

$$E\left(\frac{SCB}{b-1}\right) = \sigma^2 + \frac{k}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

y s_2^2 sobrestimará a σ^2. Un tercer estimado de σ^2, basado en $(k - 1)(b - 1)$ grados de libertad e independiente de s_1^2 y s_2^2, es

$$s^2 = \frac{SCE}{(k-1)(b-1)},$$

que es no sesgado independientemente de la veracidad o falsedad de cualquier hipótesis nula.

Para probar la hipótesis nula de que los efectos de los tratamientos son iguales a cero, se calcula la razón $f_1 = s_1^2/s^2$, que es un valor de la variable aleatoria F_1, que tiene una distribución F con $k - 1$ y $(k - 1)(b - 1)$ grados de libertad, cuando la hipótesis nula es verdadera. La hipótesis nula se rechaza al nivel de significancia α cuando

$$f_1 > f_\alpha[k-1, (k-1)(b-1)].$$

En la práctica, primero calculamos *STC*, *SCT* y *SCB,* y después, utilizando la identidad de la suma de cuadrados, obtenemos *SCE* mediante una resta. Los grados de libertad asociados con *SCE* por lo general también se obtienen por sustracción; es decir,

$$(k-1)(b-1) = kb - 1 - (k-1) - (b-1).$$

Los cálculos necesarios para un problema de análisis de varianza para un diseño de bloques completos aleatorizados se puede resumir como se observa en la tabla 13.8.

Ejemplo 13.6: Se consideran cuatro máquinas diferentes, M_1, M_2 M_3 y M_4, para ensamblar un producto específico. Se decidió que para comparar las máquinas se usarían 6 operadores distintos en un experimento de bloques aleatorizados. Las máquinas se asignaron al azar a cada operador. La operación de las máquinas requiere destreza física, y se anticipó que habría una diferencia en la velocidad con que los operadores trabajaban con las máquinas. En la tabla 13.9 se observan los tiempos (en segundos) requeridos para ensamblar el producto.

A un nivel de significancia de 0.05, pruebe la hipótesis H_0 de que las máquinas se desempeñan con el mismo índice de velocidad promedio.

Tabla 13.8: Análisis de varianza para el diseño de bloques completos aleatorizados

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Tratamientos	SCT	$k-1$	$s_1^2 = \dfrac{SCT}{k-1}$	$f_1 = \dfrac{s_1^2}{s^2}$
Bloques	SCB	$b-1$	$s_2^2 = \dfrac{SCB}{b-1}$	
Error	SCE	$(k-1)(b-1)$	$s^2 = \dfrac{SCE}{(k-1)(b-1)}$	
Total	STC	$kb-1$		

Tabla 13.9: Tiempo para ensamblar el producto, en segundos

	Operador						
Máquina	1	2	3	4	5	6	Total
1	42.5	39.3	39.6	39.9	42.9	43.6	247.8
2	39.8	40.1	40.5	42.3	42.5	43.1	248.3
3	40.2	40.5	41.3	43.4	44.9	45.1	255.4
4	41.3	42.2	43.5	44.2	45.9	42.3	259.4
Total	163.8	162.1	164.9	169.8	176.2	174.1	1010.9

***Solución*:** Las hipótesis son

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ (los efectos de las máquinas son iguales a cero),

H_1: Al menos una de las α_i no es igual a cero.

Para producir el análisis de varianza que aparece en la tabla 13.10 se emplean las fórmulas de la suma de cuadrados que se presentan en la página 536 y los grados de libertad. El valor $f = 3.34$ es significativo con $P = 0.048$. Si se emplea $\alpha = 0.05$ como al menos una aproximación burda, se concluye que las máquinas no se desempeñan con el mismo índice de velocidad media. ▟

Tabla 13.10: Análisis de varianza para los datos de la tabla 13.9

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Máquinas	15.93	3	5.31	3.34
Operadores	42.09	5	8.42	
Error	23.84	15	1.59	
Total	81.86	23		

Comentarios adicionales acerca de la formación de bloques

En el capítulo 10 presentamos un procedimiento para comparar medias cuando las observaciones estaban *ordenadas por pares*. El procedimiento implicaba "restar" el efecto

debido a la paridad homogénea para así trabajar con las diferencias. Este es un caso especial de diseño de bloques completos aleatorizados con $k = 2$ tratamientos. Las n unidades homogéneas a las que se asignaron los tratamientos adoptan el papel de bloques.

Si hay heterogeneidad en las unidades experimentales, el experimentador no debería confundirse y pensar que siempre es ventajoso reducir el error experimental mediante el uso de pequeños bloques homogéneos. De hecho podría haber casos en los que no es deseable formar bloques. El propósito de reducir la varianza del error es incrementar la *sensibilidad* de la prueba para detectar diferencias en las medias de los tratamientos. Esto se refleja en la potencia del procedimiento de prueba. (En la sección 13.11 se analiza con mayor detalle la potencia del procedimiento de prueba del análisis de varianza). La potencia para detectar ciertas diferencias entre las medias de los tratamientos se incrementa con una disminución de la varianza del error. Sin embargo, la potencia también se ve afectada por los grados de libertad con los que se estima la varianza, y la formación de bloques reduce los grados de libertad que están disponibles desde $k(b-1)$ para la clasificación de un solo factor, hasta $(k-1)(b-1)$. De modo que se podría perder potencia con la formación de bloques si no hay una reducción significativa de la varianza del error.

Interacción entre bloques y tratamientos

Otra suposición importante que está implícita en la escritura del modelo para un diseño de bloques completos aleatorizados es que los efectos de los bloques y del tratamiento son aditivos. Esto equivale a decir que

$$\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'} \quad \text{o bien} \quad \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'},$$

para cada valor de i, i', j y j'. Es decir, la diferencia entre las medias de la población para los bloques j y j' es la misma para cada tratamiento, y la diferencia entre las medias de la población para los tratamientos i e i' es la misma para cada bloque. Las líneas paralelas de la figura 13.6*a* ilustran un conjunto de respuestas medias para las cuales los efectos de los tratamientos y los bloques son aditivos, mientras que las líneas que se intersecan en la figura 13.6*b* exhiben una situación en la que se dice que los efectos de los tratamientos y de los bloques **interactúan**. Con respecto al ejemplo 13.6, si el operador 3 es en promedio 0.5 segundos más rápido que el operador 2 cuando utiliza la máquina 1, entonces el operador 3 será 0.5 segundos más rápido, en promedio, que el operador 2 cuando se empleen las máquinas 2, 3 o 4. En muchos experimentos no se cumple la suposición de aditividad y el análisis descrito en esta sección llevaría a conclusiones erróneas. Por ejemplo, suponga que el operador 3 es 0.5 segundos más rápido, en promedio, que el operador 2 si emplea la máquina 1, pero que es 0.2 segundos más lento, en promedio, que el operador 2 si utiliza la máquina 2. En ese caso los operadores y las máquinas estarían interactuando.

Una inspección de la tabla 13.9 sugiere la posible presencia de interacción. Esta aparente interacción podría ser real o podría deberse al error experimental. El análisis del ejemplo 13.6 se basó en la suposición de que la aparente interacción se debe por completo al error experimental. Si la variabilidad total de nuestros datos se debiera en parte al efecto de la interacción, esa fuente de variación seguiría formando parte de la suma de cuadrados del error, **provocando que el cuadrado medio del error sobrestime a σ^2**, incrementando así la probabilidad de cometer un error tipo II. De hecho, hemos supuesto un modelo incorrecto. Si permitimos que $(\alpha\beta)_{ij}$ denote el efecto de la interacción del

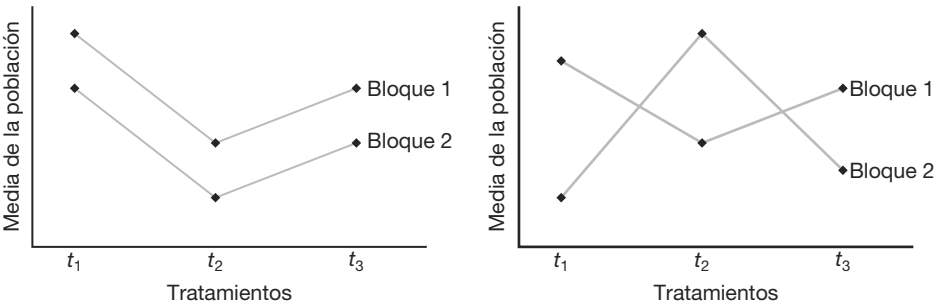


Figura 13.6: Medias de la población para *a*) resultados aditivos y *b*) efectos de interacción.

i-ésimo tratamiento y el *j*-ésimo bloque, podríamos escribir un modelo más adecuado con la forma siguiente:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij},$$

al que se impondrían las restricciones adicionales

$$\sum_{i=1}^{k} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0, \text{ para } i = 1, \ldots, k \text{ y } j = 1, \ldots, b.$$

Ahora es fácil comprobar que

$$E\left[\frac{SCE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{b} (\alpha\beta)_{ij}^2 .$$

Así, el cuadrado medio del error es considerado un **estimado sesgado de σ^2 cuando se ha ignorado la interacción existente**. En este momento parecería necesario utilizar un procedimiento para detectar la interacción en aquellos casos en que se sospecha que exista. Tal procedimiento requiere que se disponga de un estimado no sesgado e independiente de σ^2. Por desgracia, el diseño de bloques aleatorizados no se presta a una prueba de este tipo, a menos que se modifique el diseño inicial del experimento. En el capítulo 14 se estudia este tema de manera detallada.

13.9 Métodos gráficos y verificación del modelo

En varios capítulos de este libro se hace referencia a procedimientos gráficos para mostrar datos y resultados analíticos. En los primeros capítulos se usaron gráficas de tallo y hojas y de caja y extensión como auxiliares visuales para resumir muestras. En el capítulo 10 se emplearon diagnósticos similares para entender mejor los datos de dos problemas de muestreo. En el capítulo 11 se introdujo el concepto de gráfica de residuales para detectar violaciones de las suposiciones estándar. En los últimos años gran parte de la atención dedicada al análisis de datos se ha centrado en los **métodos gráficos**. Al igual que en la regresión, el análisis de varianza se presta a la elaboración de gráficas que

ayudan a resumir los datos y a detectar violaciones. Por ejemplo, una gráfica sencilla de las observaciones brutas alrededor de la media de cada tratamiento proporciona al analista una noción de la variabilidad entre las medias muestrales y dentro de las muestras. La figura 13.7 ilustra una de tales gráficas para los datos de agregados que se presentan en la tabla 13.1. A partir de la apariencia de la gráfica se obtiene incluso una idea de cuáles agregados (si los hubiera) destacan de los demás. Es evidente que el agregado 4 resalta del resto, y que los agregados 3 y 5 forman un grupo homogéneo, así como los agregados 1 y 2.

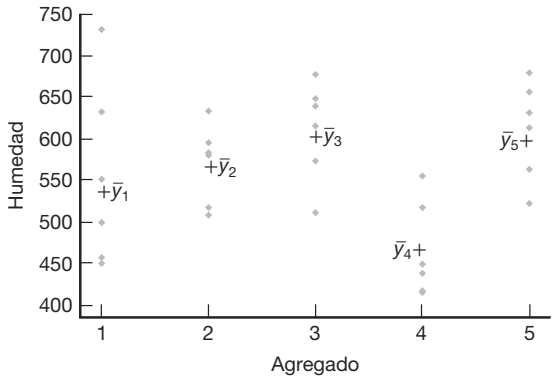


Figura 13.7: Gráfica de los datos alrededor de la media para los datos de los agregados de la tabla 13.1.

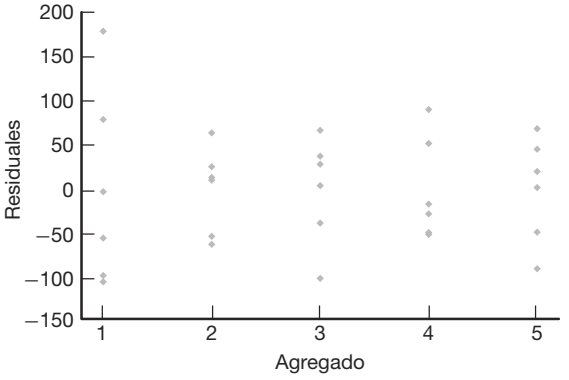


Figura 13.8: Gráfica de los residuales para cinco agregados con los datos de la tabla 13.1.

Como ocurre en el caso de la regresión, los residuales son útiles en el análisis de varianza para dar un diagnóstico sobre la detección de violaciones de los supuestos. Para formar los residuales sólo necesitamos considerar el modelo del problema de un solo factor, que es

$$y_{ij} = \mu_i + \epsilon_{ij} .$$

Es fácil determinar que el estimado de μ_i es $\bar{y}_i$. Por lo tanto, el *ij*-ésimo residual es $\bar{y}_{ij} - \bar{y}_j$, lo cual se extiende fácilmente al modelo de bloques completos aleatorizados. Sería aleccionador graficar los residuos para cada agregado con el fin de obtener cierta información sobre la suposición de varianza homogénea. Esta gráfica se muestra en la figura 13.8.

Las tendencias en gráficas como éstas podrían revelar dificultades en ciertas situaciones, especialmente cuando la violación de una suposición en particular se manifiesta en la gráfica. En el caso de la figura 13.8, los residuales parecen indicar que las varianzas *dentro de los tratamientos* son razonablemente homogéneas, excepto la del agregado 1. Hay cierta evidencia gráfica de que la varianza del agregado 1 es más grande que la del resto.

¿Qué es un residual para un diseño de BCA?

El diseño de bloques completos aleatorizados es otra situación experimental en la cual una gráfica permite que el analista se sienta cómodo con una "imagen ideal" o que tal

vez detecte dificultades. Recuerde que el modelo para el diseño de bloques completos aleatorizados es

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \qquad i = 1, \ldots, k, \qquad j = 1, \ldots, b,$$

con las restricciones impuestas

$$\sum_{i=1}^{k} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0.$$

Para determinar qué es lo que en realidad constituye un residual considere que

$$\alpha_i = \mu_{i.} - \mu, \qquad \beta_j = \mu_{.j} - \mu$$

y que μ se estima por medio de $\bar{y}..$, $\mu_{i.}$ se estima por medio de $\bar{y}_{i.}$ y $\mu_{.j}$ se estima por medio de $\bar{y}_{.j}$. Como resultado, el *valor ajustado* o pronosticado $\bar{y}_{ij}$ es dado por

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..},$$

y, entonces, el residual en la observación (i, j) es dado por

$$y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}.$$

Observe que $\bar{y}_{ij}$, el valor ajustado, es un estimado de la media μ_{ij}. Esto es congruente con la partición de la variabilidad dada en el teorema 13.3, en la que la suma de los cuadrados del error es

$$SCE = \sum_{i}^{k} \sum_{j}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

Las técnicas visuales para el diseño de bloques completos aleatorizados requieren que se grafiquen los residuos por separado para cada tratamiento y cada bloque. Si la suposición de varianza homogénea se cumple, el analista debería esperar una variabilidad aproximadamente igual. El lector seguramente recordará que en el capítulo 12 se estudiaron gráficas de los residuales con el objetivo de detectar si el modelo era inadecuado. En el caso del diseño de bloques completos aleatorizados, una grave falla del modelo podría estar relacionada con la suposición de aditividad (lo cual significa que no hay interacción). Si no hay interacción debe surgir un patrón aleatorio.

Considere los datos del ejemplo 13.6, donde los tratamientos son cuatro máquinas y los bloques son seis operadores. Las figuras 13.9 y 13.10 incluyen las gráficas de los residuales para tratamientos separados y bloques separados. La figura 13.11 presenta una gráfica de los residuales contra los valores ajustados. La figura 13.9 revela que quizá la varianza del error no sea la misma para todas las máquinas, y lo mismo podría ocurrir con la varianza del error para cada uno de los seis operadores. Sin embargo, al parecer dos residuales inusualmente grandes son los que provocan la aparente dificultad. La figura 13.11 es una gráfica de residuales que revela evidencia razonable de un comportamiento aleatorio. Sin embargo, sobresalen los dos residuales grandes ya detectados.

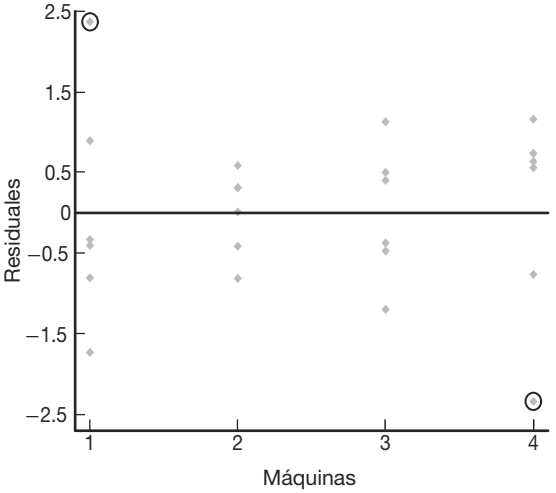


Figura 13.9: Gráfica de residuales para las cuatro máquinas de los datos del ejemplo 13.6.

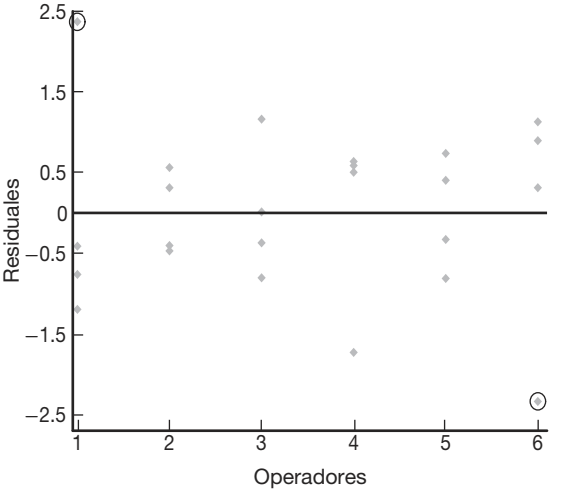


Figura 13.10: Gráfica de residuales para los seis operadores de los datos del ejemplo 13.6.

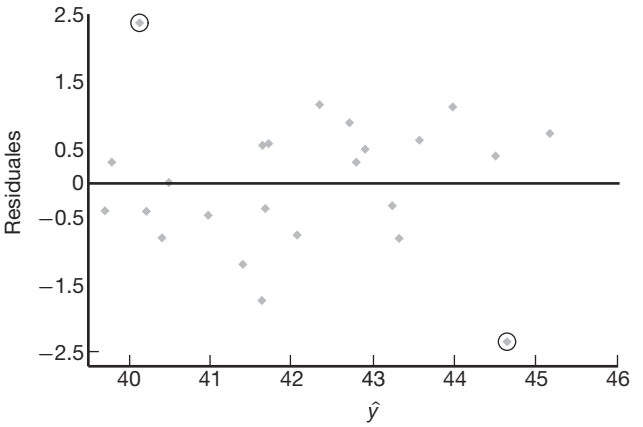


Figura 13.11: Residuales graficados contra los valores ajustados para los datos del ejemplo 13.6.

13.10 Transformaciones de datos en el análisis de varianza

En el capítulo 11 se puso mucha atención a la transformación de la respuesta y en situaciones para las que se ajustaba un modelo de regresión lineal a un conjunto de datos. Es evidente que se aplican los mismos conceptos a la regresión lineal múltiple, aunque esto no se analizó en el capítulo 12. En el estudio de los modelos de regresión se hizo énfasis en las transformaciones de y que producirían un modelo que se ajustara mejor a los datos que uno en el que la y ingresara de forma lineal. Por ejemplo, si la estructura del "tiempo" es de naturaleza exponencial, entonces una transformación logarítmica de

y linealiza la estructura y, por lo tanto, se anticipa más éxito cuando se utiliza la respuesta transformada.

Aunque el propósito fundamental de la transformación de datos que se ha analizado hasta este momento ha sido mejorar el ajuste del modelo, hay otras razones para transformar o reexpresar la respuesta y, y muchas de ellas se relacionan con las suposiciones que se hacen, por ejemplo, las suposiciones de las cuales depende la validez del análisis. Una suposición muy importante en el análisis de varianza es la de la varianza homogénea que se estudió antes en la sección 13.4. Se supone una **varianza común σ^2**. Si la varianza difiere mucho de un tratamiento a otro, y se realiza el ANOVA estándar que se estudia en este capítulo (y en otros posteriores), los resultados serían muy deficientes. En otras palabras, el análisis de varianza no es **robusto** respecto a la suposición de varianza homogénea. Como se ha dicho hasta el momento, se trata del motivo principal para la graficación de los residuales que estudiamos en la sección anterior y que ilustramos en las figuras 13.9, 13.10 y 13.11. Esas gráficas permiten detectar problemas debidos a una varianza no homogénea. Sin embargo, ¿qué hay que hacer al respecto? ¿Cómo se corrigen?

¿De dónde proviene la varianza no homogénea?

Con frecuencia, aunque no siempre, la varianza no homogénea en el ANOVA existe debido a la distribución de las respuestas. Ahora, por supuesto, se supone la normalidad de la respuesta, pero hay ciertas situaciones en las que se necesitan pruebas de las medias aunque la distribución de la respuesta sea una de las distribuciones no normales que se estudiaron en los capítulos 5 y 6, es decir, la distribución de Poisson, la logarítmica normal, la exponencial y la gamma. Realmente existen problemas del tipo del ANOVA con datos de conteo, duración antes de la falla, etcétera.

En los capítulos 5 y 6 se demostró que, además del caso de la normal, la varianza de una distribución con frecuencia será función de la media, es decir, $\sigma_i^2 = g(\mu_i)$. Por ejemplo, en el caso de la distribución de Poisson, $\text{Var}(Y_i) = \mu_i = \sigma_i^2$, lo que significa que la *varianza es igual a la media*. En el caso de la distribución exponencial, $\text{Var}(Y_i) = \sigma_i^2 = \mu_i^2$, o sea que la *varianza es igual al cuadrado de la media.* Para el caso de la logarítmica normal, una transformación logarítmica produce una distribución normal con varianza constante σ^2.

Los mismos conceptos que usamos en el capítulo 4 para determinar la varianza de una función no lineal pueden ayudarnos a determinar la naturaleza de la *transformación estabilizadora de la varianza* $g(y_i)$. Recuerde la expansión de las series de Taylor de primer orden de $g(y_i)$ alrededor de $y_i = \mu_i$, donde $g'(\mu_i) = \left[\dfrac{\partial g(y_i)}{\partial y_i}\right]_{y_i = \mu_i}$. La función de transformación $g(y)$ debe ser independiente de μ para que baste como la transformación estabilizadora de la varianza. De lo anterior

$$\text{Var}[g(y_i)] \approx [g'(\mu_i)]^2 \sigma_i^2.$$

Como resultado, $g(y_i)$ debe ser tal que $g'(\mu_i) \propto \frac{1}{\sigma}$. Así, si se sospecha que la respuesta tiene una distribución de Poisson, $\sigma_i = \mu_i^{1/2}$, de modo que $g'(\mu_i) \propto \frac{1}{\mu_i^{1/2}}$. Entonces, la transformación estabilizadora de la varianza es $g(y_i) = y_i^{1/2}$. A partir de esta ilustración y manipulaciones similares para las distribuciones exponencial y gamma, se obtiene lo siguiente.

Distribución	Transformaciones estabilizadoras de la varianza
Poisson	$g(y) = y^{1/2}$
Exponencial	$g(y) = \ln y$
Gamma	$g(y) = \ln y$

Ejercicios

13.25 Se utilizaron cuatro clases de fertilizante f_1, f_2, f_3 y f_4 para estudiar la cosecha de frijol. El suelo se dividió en 3 bloques, cada uno de los cuales contiene 4 parcelas homogéneas. A continuación se presentan las cosechas en kilogramos por parcela, así como los tratamientos correspondientes:

Bloque 1	Bloque 2	Bloque 3
$f_1 = 42.7$	$f_3 = 50.9$	$f_4 = 51.1$
$f_3 = 48.5$	$f_1 = 50.0$	$f_2 = 46.3$
$f_4 = 32.8$	$f_2 = 38.0$	$f_1 = 51.9$
$f_2 = 39.3$	$f_4 = 40.2$	$f_3 = 53.5$

Realice un análisis de varianza a un nivel de significancia de 0.05 utilizando el modelo de bloques completos aleatorizados.

13.26 Se compararon las cosechas de tres variedades de papas. El experimento se efectuó asignando cada variedad de manera aleatoria a 3 parcelas del mismo tamaño, en 4 lugares diferentes. Se registraron las siguientes cosechas para las variedades *A*, *B* y *C*, en 100 kilogramos por parcela:

Lugar 1	Lugar 2	Lugar 3	Lugar 4
B: 13	*C*: 21	*C*: 9	*A*: 11
A: 18	*A*: 20	*B*: 12	*C*: 10
C: 12	*B*: 23	*A*: 14	*B*: 17

Realice un análisis de varianza de bloques completos aleatorizados con el objetivo de probar la hipótesis de que no hay diferencia en la capacidad de rendimiento de las 3 variedades de papas. Utilice un nivel de significancia de 0.05 y saque conclusiones.

13.27 Los siguientes datos son los porcentajes de aditivos extranjeros, medidos por 5 analistas, de 3 marcas distintas de mermelada de fresa, *A*, *B* y *C*.

Analista 1	Analista 2	Analista 3	Analista 4	Analista 5
B: 2.7	*C*: 7.5	*B*: 2.8	*A*: 1.7	*C*: 8.1
C: 3.6	*A*: 1.6	*A*: 2.7	*B*: 1.9	*A*: 2.0
A: 3.8	*B*: 5.2	*C*: 6.4	*C*: 2.6	*B*: 4.8

A un nivel de significancia de 0.05, realice un análisis de varianza de bloques completos aleatorizados para probar la hipótesis de que el porcentaje de aditivos extranjeros es el mismo para las tres marcas de mermelada. ¿Cuál de ellas parece tener menos aditivos?

13.28 Los siguientes datos representan las calificaciones finales obtenidas por 5 estudiantes en matemáticas, inglés, francés y biología:

	Materia			
Estudiante	**Matemáticas**	**Inglés**	**Francés**	**Biología**
1	68	57	73	61
2	83	94	91	86
3	72	81	63	59
4	55	73	77	66
5	92	68	75	87

Pruebe la hipótesis de que los cursos tienen la misma dificultad. Use un valor *P* en sus conclusiones y analice sus hallazgos.

13.29 En el estudio *The Periphyton of the South River, Virginia: Mercury Concentration, Productivity, and Autotropic Index Studies*, efectuado por el Departamento de Ciencias e Ingeniería Ambientales de Virginia Tech, se midió la concentración total de mercurio en sólidos totales de perifitón en seis estaciones distintas durante seis días diferentes. Determine si el contenido medio de mercurio difiere significativamente entre las estaciones utilizando los siguientes datos. Use un valor *P* y analice sus hallazgos.

	Estación					
Fecha	***CA***	***CB***	***E*1**	***E*2**	***E*3**	***E*4**
8 de abril	0.45	3.24	1.33	2.04	3.93	5.93
23 de junio	0.10	0.10	0.99	4.31	9.92	6.49
1 de julio	0.25	0.25	1.65	3.13	7.39	4.43
8 de julio	0.09	0.06	0.92	3.66	7.88	6.24
15 de julio	0.15	0.16	2.17	3.50	8.82	5.39
23 de julio	0.17	0.39	4.30	2.91	5.50	4.29

13.30 Una planta de energía nuclear produce una gran cantidad de calor que generalmente se descarga en los sistemas de agua. Ese calor eleva la temperatura del líquido, lo cual da como resultado una mayor concentración de clorofila *a* que, a su vez, alarga la temporada de crecimiento. Para estudiar este efecto se tomaron muestras de agua mensualmente en 3 estaciones, durante un periodo de 12 meses. La estación *A* es la que se ubica más cerca de una descarga potencial de agua caliente, la estación *C* es la más lejana de la descarga y la estación *B* se encuentra entre las estaciones *A* y *C*. Se registraron las siguientes concentraciones de clorofila *a*.

	Estación		
Mes	***A***	***B***	***C***
Enero	9.867	3.723	4.410
Febrero	14.035	8.416	11.100
Marzo	10.700	20.723	4.470
Abril	13.853	9.168	8.010
Mayo	7.067	4.778	34.080
Junio	11.670	9.145	8.990
Julio	7.357	8.463	3.350
Agosto	3.358	4.086	4.500
Septiembre	4.210	4.233	6.830
Octubre	3.630	2.320	5.800
Noviembre	2.953	3.843	3.480
Diciembre	2.640	3.610	3.020

Realice un análisis de varianza y, a un nivel de significancia de 0.05, pruebe la hipótesis de que no hay diferencia en las concentraciones medias de clorofila *a* en las 3 estaciones.

13.31 En un estudio realizado por el Departamento de Salud y Educación Física de Virginia Tech, se asignaron 3 dietas durante 3 días a 6 sujetos utilizando un diseño de bloques completos aleatorizados. Los sujetos, que desempeñan el papel de bloques, recibieron las siguientes 3 dietas en orden aleatorio:

Dieta 1: grasas mixtas y carbohidratos,
Dieta 2: alta en grasas,
Dieta 3: alta en carbohidratos.

Al terminar el periodo de tres días se puso a cada sujeto en una banda caminadora y se midió el tiempo, en segundos, que transcurría hasta que se sentían exhaustos. Efectúe un análisis de varianza separando la dieta, los sujetos y la suma de cuadrados del error. Utilice un valor *P* para determinar si existen diferencias significativas entre las dietas. Los datos registrados son los siguientes:

	Sujeto					
Dieta	**1**	**2**	**3**	**4**	**5**	**6**
1	84	35	91	57	56	45
2	91	48	71	45	61	61
3	122	53	110	71	91	122

13.32 El personal forestal utiliza arsénico orgánico como arboricida. La cantidad de arsénico que absorbe el cuerpo cuando se expone a este producto constituye un grave problema de salud. Es importante que la cantidad de exposición se determine rápido, de manera que pueda retirarse del trabajo a los empleados con niveles elevados de arsénico. En un experimento descrito en el artículo "A Rapid Method for the Determination of Arsenic Concentrations in Urine at Field Locations", publicado en el *American Industrial Hygiene Association Journal* (Vol. 37, 1976), especímenes de orina de 4 personas del servicio forestal fueron divididos por igual en tres muestras para que pudiera analizarse el contenido de arsénico en la orina de cada individuo en un laboratorio universitario: las muestras eran analizadas por un químico con un sistema portátil, así como también por un empleado forestal que había recibido una capacitación breve. Se registraron los siguientes niveles de arsénico, en partes por millón:

	Analista		
Individuo	**Empleado**	**Químico**	**Laboratorio**
1	0.05	0.05	0.04
2	0.05	0.05	0.04
3	0.04	0.04	0.03
4	0.15	0.17	0.10

Realice un análisis de varianza y, a un nivel de significancia de 0.05, pruebe la hipótesis de que no hay diferencia en los niveles de arsénico con los tres métodos de análisis.

13.33 Los científicos del Departamento de Patología Vegetal de Virginia Tech realizaron un experimento en el que se aplicaron 5 tratamientos diferentes en 6 lugares distintos de un huerto de manzanas para determinar si había diferencias significativas en el crecimiento entre los tratamientos. Los tratamientos 1 a 4 representan distintos herbicidas y el 5 es un control. El periodo de crecimiento fue de mayo a noviembre de 1982, y los datos de crecimiento nuevo, medido en centímetros, para muestras seleccionadas de los 6 lugares en el huerto, son los siguientes:

	Ubicaciones					
Tratamiento	**1**	**2**	**3**	**4**	**5**	**6**
1	455	72	61	215	695	501
2	622	82	444	170	437	134
3	695	56	50	443	701	373
4	607	650	493	257	490	262
5	388	263	185	103	518	622

Lleve a cabo un análisis de varianza, separando el tratamiento, el lugar y la suma de cuadrados del error. Determine si hay diferencias significativas entre las medias de los tratamientos. Mencione un valor *P*.

13.34 En el artículo "Self-Control and Therapist Control in the Behavioral Treatment of Overweight Women", publicado en *Behavioral Research and Therapy* (Vol. 10, 1972), se estudiaron dos tratamientos de reducción y otro de control para observar sus efectos en el cambio del peso en mujeres obesas. Los dos tratamientos reductores involucrados fueron un programa autodirigido de reducción de peso y otro controlado por un terapeuta. Se asignó a cada uno de 10 sujetos a uno de los 3 programas de tratamiento en orden aleatorio y se midió la pérdida de peso. Se registraron los siguientes cambios en el peso:

	Tratamiento		
Sujeto	**Control**	**Autodirigido**	**Con terapeuta**
1	1.00	−2.25	−10.50
2	3.75	−6.00	−13.50
3	0.00	−2.00	0.75
4	−0.25	−1.50	−4.50
5	−2.25	−3.25	−6.00
6	−1.00	−1.50	4.00
7	−1.00	−10.75	−12.25
8	3.75	−0.75	−2.75
9	1.50	0.00	−6.75
10	0.50	−3.75	−7.00

Realice un análisis de varianza y, a un nivel de significancia de 0.01, pruebe la hipótesis de que no hay diferencia en las pérdidas de peso promedio para los 3 tratamientos. ¿Cuál tratamiento fue el mejor?

13.35 En el libro *Design of Experiments for the Quality Improvement*, publicado por la Japanese Standards Association (1989) se reportó un estudio sobre la cantidad de tinta que se requiere para obtener el mejor color para cierto tipo de tela. En dos plantas diferentes se administraron tres cantidades de tinta: $\frac{1}{3}$ del porcentaje de wof, es decir, $\frac{1}{3}$ del porcentaje del peso de la tela, 1% de wof y 3% de wof. Después se observó cuatro veces la densidad del color de la tela para cada nivel de tinta aplicada en cada planta.

	Cantidad de tinta					
	1/3%		**1%**		**3%**	
Planta 1	5.2	6.0	12.3	10.5	22.4	17.8
	5.9	5.9	12.4	10.9	22.5	18.4
Planta 2	6.5	5.5	14.5	11.8	29.0	23.2
	6.4	5.9	16.0	13.6	29.7	24.0

A un nivel de significancia de 0.05, realice un análisis de varianza para probar la hipótesis de que no hay diferencia en la densidad de color de la tela con los tres niveles de tinta. Considere a las plantas como bloques.

13.36 Se realizó un experimento con el fin de comparar tres tipos de materiales para recubrir alambres de cobre. El propósito del recubrimiento consiste en eliminar los "defectos" del alambre. A cada recubrimiento se le asignaron al azar 10 especímenes distintos, de 5 milímetros de longitud, para que les fuera aplicado. Después se sometió a los 30 especímenes a cierto tipo de desgaste abrasivo. Al final se midió el número de defectos en cada uno y se obtuvieron los siguientes resultados:

Material												
1				**2**				**3**				
6	8	4	5	3	3	5	4	12	8	7	14	
7	7	9	6	2	4	4	5	18	6	7	18	
7	8			4	3			8	5			

Suponga que se acepta que se puede aplicar un proceso de Poisson, por lo que el modelo es $Y_{ij} = \mu_i + \epsilon_{ij}$, donde μ_i es la media de la distribución de Poisson y $\sigma^2_{Y_{ij}} = \mu_i$.

a) Haga una transformación apropiada de los datos y un análisis de varianza.
b) Determine si hay evidencia suficiente para preferir un material de recubrimiento sobre los demás. Muestre cualesquiera hallazgos que sugieran una conclusión.
c) Haga una gráfica de residuales y coméntela.
d) Mencione el propósito de la transformación de los datos.
e) ¿Qué otra suposición se hace en este caso, que quizá la transformación no cumpla por completo?
f) Comente el inciso *e* después de elaborar una gráfica de probabilidad normal sobre los residuales.

13.11 Modelos de efectos aleatorios

A lo largo de este capítulo estudiamos los procedimientos del análisis de varianza en los que el objetivo principal es estudiar el efecto sobre ciertas respuestas de ciertos tratamientos fijos o predeterminados. Los experimentos en los que los tratamientos o los niveles de tratamiento son preseleccionados por el experimentador, y no elegidos al azar, se denominan **experimentos de efectos fijos**. Para el modelo de efectos fijos sólo se hacen inferencias acerca de los tratamientos específicos que se utilizaron en el experimento.

Con frecuencia es importante que el experimentador sea capaz de hacer inferencias acerca de una población de tratamientos a través de un experimento en el que los tratamientos empleados se elijan al azar de entre la población. Por ejemplo, un biólogo podría estar interesado en saber si hay o no una varianza significativa en alguna característica fisiológica debida a un tipo de animal. Los tipos de animales que en realidad se usan en el experimento se eligen al azar y representan los efectos del tratamiento. Un químico podría estar interesado en estudiar el efecto de los laboratorios sobre el análisis químico de

una sustancia; no le interesa un laboratorio en particular, sino una población grande de laboratorios. Así, podría seleccionar al azar un grupo de laboratorios y asignar muestras a cada uno para su análisis. Entonces, la inferencia estadística implicaría 1) probar si los laboratorios contribuyen o no a una varianza diferente de cero en los resultados de los análisis, y 2) estimar la varianza debida a los laboratorios y a la varianza dentro de los mismos.

Modelo y suposiciones para el modelo de efectos aleatorios

El **modelo de efectos aleatorios** de un solo factor se escribe como el modelo de efectos fijos, pero sus términos tienen significados diferentes. La respuesta $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ es ahora un valor de la variable aleatoria

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \text{ con } i = 1, 2, \ldots, k \text{ y } j = 1, 2, \ldots, n,$$

donde las A_i tienen distribución normal e independiente con media igual a cero y varianza σ_α^2, y son independientes de las ϵ_{ij}. Al igual que para el modelo de efectos fijos, las ϵ_{ij} también tienen distribución normal e independiente con media igual a cero y varianza σ^2. Observe que para un experimento de efectos aleatorios, ya no se aplica la restricción de que $\sum_{i=1}^{k} \alpha_i = 0$.

Teorema 13.4: Para el modelo del análisis de varianza de efectos aleatorios de un solo factor,

$$E(SCT) = (k-1)\sigma^2 + n(k-1)\sigma_\alpha^2 \qquad \text{y} \qquad E(SCE) = k(n-1)\sigma^2.$$

La tabla 13.11 presenta los cuadrados medios esperados tanto para un experimento de efectos fijos como para uno de efectos aleatorios. Los cálculos para un experimento de efectos aleatorios se realizan exactamente de la misma forma que los del experimento de efectos fijos. Es decir, la suma de cuadrados, los grados de libertad y las columnas de los cuadrados medios en la tabla del análisis de varianza son iguales para ambos modelos.

Tabla 13.11: Cuadrados medios esperados para el experimento de un solo factor

Fuente de variación	Grados de libertad	Cuadrados medios	Cuadrados medios esperados	
			Efectos fijos	Efectos aleatorios
Tratamientos	$k-1$	s_1^2	$\sigma^2 + \frac{n}{k-1}\sum_i \alpha_i^2$	$\sigma^2 + n\sigma_\alpha^2$
Error	$k(n-1)$	s^2	σ^2	σ^2
Total	$nk-1$			

Para el modelo de efectos aleatorios la hipótesis de que todos los efectos del tratamiento son iguales a cero se escribe como sigue:

Hipótesis para un experimento de efectos aleatorios

$$H_0: \sigma_\alpha^2 = 0,$$
$$H_1: \sigma_\alpha^2 \neq 0.$$

Esta hipótesis afirma que los diferentes tratamientos no contribuyen en absoluto a la variabilidad de la respuesta. De la tabla 13.11 es evidente que tanto s_1^2 como s^2 son estimados de σ^2 cuando H_0 es verdadera, y que la razón

$$f = \frac{s_1^2}{s^2}$$

es un valor de la variable aleatoria F que tiene la distribución F con $k-1$ y $k(n-1)$ grados de libertad. La hipótesis nula se rechaza a un nivel de significancia α cuando

$$f > f_\alpha[k-1, k(n-1)].$$

En muchos estudios científicos y de ingeniería el interés no se centra en la prueba F. El científico sabe que el efecto aleatorio, en efecto, es significativo. Lo más importante es la estimación de los diversos componentes de la varianza. Esto produce un sentido de *jerarquía* en términos de cuáles factores producen la mayor variabilidad y en qué cantidad. En este contexto podría ser interesante cuantificar cuánto más grande es el *componente de la varianza de un solo factor* que el producido por el azar (variación aleatoria).

Estimación de los componentes de la varianza

La tabla 13.11 también se utiliza para estimar los **componentes de la varianza** σ^2 y σ_α^2. Como s_1^2 estima $\sigma^2 + n\sigma_\alpha^2$, y s^2 estima σ^2,

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s^2}{n}.$$

Ejemplo 13.7: Los datos de la tabla 13.12 representan observaciones codificadas sobre el producto de un proceso químico en el que se utilizan 5 lotes de materia prima seleccionados al azar. Demuestre que el componente de la varianza del lote es significativamente mayor que cero y obtenga su estimado.

Tabla 13.12: Datos para el ejemplo 13.7

Lote	**1**	**2**	**3**	**4**	**5**	
	9.7	10.4	15.9	8.6	9.7	
	5.6	9.6	14.4	11.1	12.8	
	8.4	7.3	8.3	10.7	8.7	
	7.9	6.8	12.8	7.6	13.4	
	8.2	8.8	7.9	6.4	8.3	
	7.7	9.2	11.6	5.9	11.7	
	8.1	7.6	9.8	8.1	10.7	
Total	55.6	59.7	80.7	58.4	75.3	329.7

***Solución*:** La suma total de cuadrados, la del lote y la suma de cuadrados del error son, respectivamente,

$$STC = 194.64,\ SCT = 72.60 \text{ y } SCE = 194.64 - 72.60 = 122.04.$$

En la tabla 13.13 se presentan estos resultados con el resto de los cálculos.

Tabla 13.13: Análisis de la varianza para el ejemplo 13.7

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Lotes	72.60	4	18.15	4.46
Error	122.04	30	4.07	
Total	194.64	34		

La razón f es significativa al nivel $\alpha = 0.05$, lo que indica que se rechaza la hipótesis de un componente del lote igual a cero. Una estimación del componente de la varianza del lote es

$$\hat{\sigma}_{\alpha}^2 = \frac{18.15 - 4.07}{7} = 2.01.$$

Observe que mientras que el **componente de la varianza del lote** es significativamente diferente de cero, cuando se compara contra el estimado de σ^2, es decir, $\hat{\sigma}^2 = CME = 4.07$, pareciera que el componente de varianza del lote no es considerablemente grande.

Si el resultado que se obtiene con la fórmula para σ_{α}^2 es negativo, o sea, que s_1^2 es menor que s^2, entonces a $\hat{\sigma}_{\alpha}^2$ se le asigna un valor de cero. Éste es un estimador sesgado. Para tener un mejor estimador de σ_{α}^2, suele emplearse un método llamado **verosimilitud restringida** (o **residual**) **máxima (REML** por sus siglas en inglés) (véase Harville, 1977, en la bibliografía). Este tipo de estimador se puede encontrar en muchos paquetes estadísticos para computadora. Los detalles de dicho procedimiento rebasan el alcance de este libro.

Diseño de bloques aleatorizados con bloques aleatorios

En un experimento de bloques completos aleatorizados, donde los bloques representan días, es concebible que el experimentador quiera que los resultados se apliquen no sólo a los días reales utilizados en el análisis, sino a cada día del año. Entonces, seleccionaría al azar los días en que se haría el experimento, así como los tratamientos y el modelo de efectos aleatorios a utilizar.

$$Y_{ij} = \mu + A_i + B_j + \epsilon_{ij}, \text{ para } i = 1, 2, \ldots, k \text{ y } j = 1, 2, \ldots, b,$$

donde las A_i, B_j y ϵ_{ij} son variables aleatorias independientes con medias igual a cero y varianzas σ_{α}^2, σ_{β}^2 y σ^2, respectivamente. Se obtienen los cuadrados medios esperados para un diseño de bloques completos aleatorizados de efectos aleatorios usando el mismo procedimiento que se usó en el problema de un solo factor; en la tabla 13.14 se presentan junto con los de un experimento de efectos fijos.

Nuevamente, los cálculos para las sumas de cuadrados y grados de libertad individuales son idénticos a los del modelo de efectos fijos. Las hipótesis

$$H_0: \ \sigma_{\alpha}^2 = 0,$$
$$H_1: \ \sigma_{\alpha}^2 \neq 0$$

se obtienen calculando

$$f = \frac{s_1^2}{s^2}$$

Tabla 13.14: Cuadrados medios esperados para un diseño de bloques completos aleatorizados

Fuente de variación	Grados de libertad	Cuadrados medios	Cuadrados medios esperados	
			Efectos fijos	Efectos aleatorios
Tratamientos	$k-1$	s_1^2	$\sigma^2 + \frac{b}{k-1}\sum_i \alpha_i^2$	$\sigma^2 + b\sigma_\alpha^2$
Bloques	$b-1$	s_2^2	$\sigma^2 + \frac{k}{b-1}\sum_j \beta_j^2$	$\sigma^2 + k\sigma_\beta^2$
Error	$(k-1)(b-1)$	s^2	σ^2	σ^2
Total	$kb-1$			

y H_0 se rechaza cuando $f > f_\alpha[k-1, (b-1)(k-1)]$.

Los estimados no sesgados de los componentes de la varianza son

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s^2}{b}, \qquad \hat{\sigma}_\beta^2 = \frac{s_2^2 - s^2}{k}.$$

Las pruebas de las hipótesis referentes a los diversos componentes de la varianza se realizan calculando las razones de los cuadrados medios adecuados, tal como se indica en la tabla 13.14, y comparándolos con los valores *f* correspondientes de la tabla A.6.

13.12 Estudio de caso

Estudio de caso 13.1: **Análisis químico**. Se pidió al personal del Departamento de Química de Virginia Tech que analizara un conjunto de datos que se obtuvo para comparar 4 métodos distintos de análisis del aluminio en cierta mezcla deflagradora sólida. Para considerar una amplia gama de laboratorios de análisis se utilizaron 5 de ellos en el experimento. Se seleccionaron esos laboratorios porque suelen realizar esa clase de análisis. Se asignaron al azar 20 muestras de material deflagrador que contenían 2.70% de aluminio, cuatro a cada laboratorio, y se dieron instrucciones acerca de cómo efectuar los análisis químicos utilizando los cuatro métodos. Los datos que se obtuvieron son los siguientes:

	Laboratorio					
Método	1	2	3	4	5	Media
A	2.67	2.69	2.62	2.66	2.70	2.668
B	2.71	2.74	2.69	2.70	2.77	2.722
C	2.76	2.76	2.70	2.76	2.81	2.758
D	2.65	2.69	2.60	2.64	2.73	2.662

Los laboratorios no se consideran efectos aleatorios, ya que no fueron seleccionados al azar de entre una población más grande de ellos. Los datos se analizaron como un diseño de bloques completos aleatorizados. Se dibujaron gráficas de los datos para determinar si era apropiado un modelo aditivo del tipo:

$$y_{ij} = \mu + m_i + l_j + \epsilon_{ij}$$

en otras palabras, un modelo con efectos aditivos. El bloque aleatorizado no es adecuado cuando existe interacción entre los laboratorios y los métodos. Considere la gráfica de la figura 13.12. Aunque es un poco difícil de interpretar porque cada punto representa una sola observación, parece que no hay interacción evidente entre los métodos y los laboratorios.

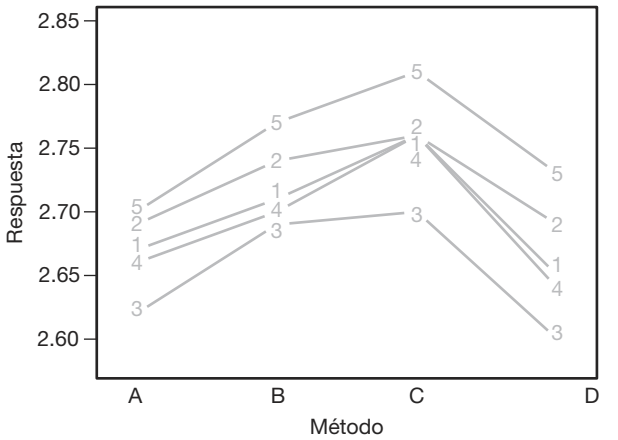


Figura 13.12: Gráfica de interacción para los datos del estudio de caso 13.1.

Gráficas de residuales

Las gráficas de residuales se usaron como indicaciones de diagnóstico con respecto a la suposición de una varianza homogénea. La figura 13.13 presenta una gráfica de residuales contra los métodos de análisis. La variabilidad descrita en los residuales parece ser bastante homogénea. Para completar, en la figura 13.14 se presenta una gráfica de probabilidad normal de los residuales.

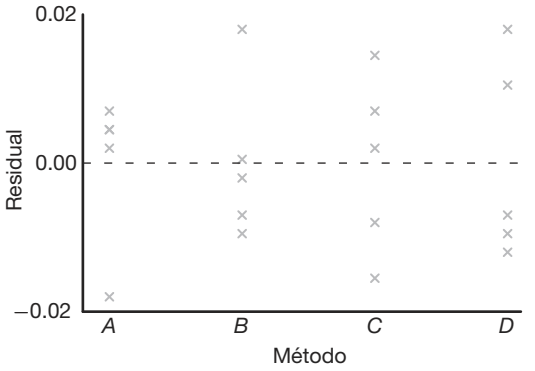


Figura 13.13: Gráfica de residuales en comparación con el método para los datos del estudio de caso 13.1.

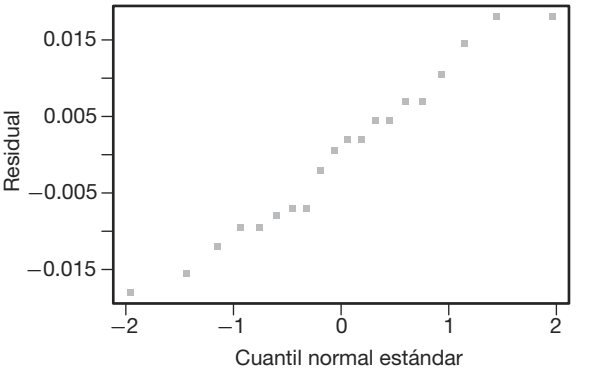


Figura 13.14: Gráfica de probabilidad normal de residuales para los datos del estudio de caso 13.1.

Las gráficas de residuales no muestran problemas con la suposición de errores normales ni con la de varianza homogénea. Para hacer el análisis de varianza se utilizó la

función PROC GLM del programa *SAS*. En la figura 13.15 se incluye una salida de resultados por computadora con comentarios.

Los valores f y P calculados sí indican una diferencia significativa entre los métodos de análisis. A este análisis le puede seguir un análisis de comparación múltiple para determinar en dónde están las diferencias entre los métodos.

Ejercicios

13.37 Al probar muestras de sangre de un paciente para detectar anticuerpos del VIH un espectrómetro determina la densidad óptica de cada muestra. La densidad óptica se mide como la absorbencia de la luz de cierta longitud de onda. La muestra de sangre es positiva si excede a cierto valor límite que se determina con muestras de control para esa corrida. A los investigadores les interesa comparar la variabilidad del laboratorio para los valores de control positivo. Los datos representan valores de control positivo para 10 corridas distintas en cuatro laboratorios seleccionados al azar.

	Laboratorio			
Corrida	**1**	**2**	**3**	**4**
1	0.888	1.065	1.325	1.232
2	0.983	1.226	1.069	1.127
3	1.047	1.332	1.219	1.051
4	1.087	0.958	0.958	0.897
5	1.125	0.816	0.819	1.222
6	0.997	1.015	1.140	1.125
7	1.025	1.071	1.222	0.990
8	0.969	0.905	0.995	0.875
9	0.898	1.140	0.928	0.930
10	1.018	1.051	1.322	0.775

a) Escriba un modelo adecuado para este experimento.
b) Estime el componente de varianza del laboratorio y la varianza dentro de los laboratorios.

13.38 Se efectúa un experimento en el que se compararán 4 tratamientos en 5 bloques. Los datos son los siguientes:

	Bloque				
Tratamiento	**1**	**2**	**3**	**4**	**5**
1	12.8	10.6	11.7	10.7	11.0
2	11.7	14.2	11.8	9.9	13.8
3	11.5	14.7	13.6	10.7	15.9
4	12.6	16.5	15.4	9.6	17.1

a) Suponga que se trata de un modelo de efectos aleatorios y pruebe la hipótesis de que no hay diferencia entre las medias de los tratamientos, a un nivel de significancia de 0.05.
b) Calcule estimados de los componentes de la varianza del tratamiento y del bloque.

13.39 Los siguientes datos muestran el efecto de cuatro operadores, elegidos al azar, sobre la producción de una máquina específica:

Operador			
1	**2**	**3**	**4**
175.4	168.5	170.1	175.2
171.7	162.7	173.4	175.7
173.0	165.0	175.7	180.1
170.5	164.1	170.7	183.7

a) Realice un análisis de varianza de efectos aleatorios a un nivel de significancia de 0.05.
b) Calcule un estimado del componente de la varianza del operador y del componente de la varianza del error experimental.

13.40 De cinco "vaciados" de metales se tomaron cinco muestras del núcleo y en cada una se analizó la cantidad de un elemento traza**.** Los siguientes son los datos de los 5 vaciados seleccionados al azar:

	Vaciado				
Núcleo	**1**	**2**	**3**	**4**	**5**
1	0.98	0.85	1.12	1.21	1.00
2	1.02	0.92	1.68	1.19	1.21
3	1.57	1.16	0.99	1.32	0.93
4	1.25	1.43	1.26	1.08	0.86
5	1.16	0.99	1.05	0.94	1.41

a) La intención es que los vaciados sean idénticos. Por lo tanto, pruebe que el componente de la varianza del "vaciado" es igual a cero. Saque conclusiones.
b) Realice un ANOVA completo y obtenga un estimado de la varianza dentro del vaciado.

13.41 Una empresa textil produce cierta tela en un número grande de telares. Los gerentes quieren que los telares sean homogéneos para que la tela que producen tenga una resistencia uniforme. Se sospecha que puede haber una variación significativa entre la resistencia de los telares. Considere los siguientes datos para 4 telares seleccionados al azar. Cada observación es una determinación de la resistencia de la tela expresada en libras por pulgada cuadrada.

Telar			
1	**2**	**3**	**4**
99	97	94	93
97	96	95	94
97	92	90	90
96	98	92	92

a) Escriba un modelo para el experimento.
b) ¿El componente de la varianza del telar difiere significativamente de cero?
c) Haga comentarios sobre la sospecha de los gerentes.

```
                          The GLM Procedure
                      Class Level Information
                Class          Levels     Values
                Method              4     A B C D
                Lab                 5     1 2 3 4 5
              Number of Observations Read          20
              Number of Observations Used          20
Dependent Variable: Response
                          Sum of
Source            DF      Squares    Mean Square   F Value    Pr > F
Model              7   0.05340500     0.00762929     42.19    <.0001
Error             12   0.00217000     0.00018083
Corrected Total   19   0.05557500

R-Square     Coeff Var       Root MSE     Response Mean
0.960954      0.497592       0.013447          2.702500

Source            DF   Type III SS    Mean Square   F Value     Pr > F
Method             3     0.03145500     0.01048500     57.98     <.0001
Lab                4     0.02195000     0.00548750     30.35     <.0001

Observation         Observed           Predicted            Residual
          1      2.67000000          2.66300000          0.00700000
          2      2.71000000          2.71700000         -0.00700000
          3      2.76000000          2.75300000          0.00700000
          4      2.65000000          2.65700000         -0.00700000
          5      2.69000000          2.68550000          0.00450000
          6      2.74000000          2.73950000          0.00050000
          7      2.76000000          2.77550000         -0.01550000
          8      2.69000000          2.67950000          0.01050000
          9      2.62000000          2.61800000          0.00200000
         10      2.69000000          2.67200000          0.01800000
         11      2.70000000          2.70800000         -0.00800000
         12      2.60000000          2.61200000         -0.01200000
         13      2.66000000          2.65550000          0.00450000
         14      2.70000000          2.70950000         -0.00950000
         15      2.76000000          2.74550000          0.01450000
         16      2.64000000          2.64950000         -0.00950000
         17      2.70000000          2.71800000         -0.01800000
         18      2.77000000          2.77200000         -0.00200000
         19      2.81000000          2.80800000          0.00200000
         20      2.73000000          2.71200000          0.01800000
```

Figura 13.15: Salida de resultados por computadora del *SAS* para los datos del estudio de caso 13.1.

Ejercicios de repaso

13.42 El Centro de Consultoría en Estadística de Virginia Tech, junto con el Departamento de Silvicultura, llevaron a cabo un análisis. Se aplicó cierto tratamiento a tres cepas de árbol. Se empleó el producto químico Garlon con el fin de regenerar las raíces de las cepas. Se usó un aerosol con cuatro niveles de concentración de Garlon. Después de cierto tiempo, se observó la altura de los retoños. Realice un análisis de varianza de un solo factor con los siguientes datos. Haga pruebas para saber si la concentración de Garlon tiene un efecto significativo sobre la altura de los retoños. Emplee $\alpha = 0.05$.

Nivel de Garlon							
1		**2**		**3**		**4**	
2.87	2.31	3.27	2.66	2.39	1.91	3.05	0.91
3.91	2.04	3.15	2.00	2.89	1.89	2.43	0.01

13.43 Considere los datos de los agregados del ejemplo 13.1. Efectúe una prueba de Bartlett a un nivel de significancia $\alpha = 0.1$ para determinar si hay heterogeneidad en la varianza entre los agregados.

13.44 En un proceso químico se utilizaron 3 catalizadores y también se incluyó un control (no catalizador). Se tienen los datos siguientes de la producción del proceso:

	Catalizador		
Control	**1**	**2**	**3**
74.5	77.5	81.5	78.1
76.1	82.0	82.3	80.2
75.9	80.6	81.4	81.5
78.1	84.9	79.5	83.0
76.2	81.0	83.0	82.1

Use una prueba de Dunnett a un nivel de significancia $\alpha = 0.01$ para determinar si se obtuvo una producción significativamente más alta con los catalizadores que sin ellos.

13.45 Se emplean cuatro laboratorios para efectuar análisis químicos. Se envían muestras del mismo material a los laboratorios para que, como parte del estudio, las analicen para determinar si dan o no, en promedio, los mismos resultados. Los resultados analíticos de los cuatro laboratorios son los siguientes:

Laboratorio			
A	***B***	***C***	***D***
58.7	62.7	55.9	60.7
61.4	64.5	56.1	60.3
60.9	63.1	57.3	60.9
59.1	59.2	55.2	61.4
58.2	60.3	58.1	62.3

a) Utilice una prueba de Bartlett para demostrar que las varianzas dentro de los laboratorios no difieren de manera significativa a un nivel de significancia $\alpha = 0.05$.

b) Realice el análisis de varianza y saque conclusiones acerca de los laboratorios.

c) Dibuje una gráfica de probabilidad normal de residuales.

13.46 Se diseñó un experimento para el personal del Departamento de Ciencia Animal de Virginia Tech, con el propósito de estudiar el tratamiento con urea y amoniaco acuoso de la espiga del trigo. El propósito era mejorar el valor nutricional para las ovejas macho. Los tratamientos dietéticos son: control, urea en la alimentación, espiga tratada con amoniaco, espiga tratada con urea. En el experimento se emplearon 24 ovejas y se separaron de acuerdo con su peso relativo. En cada grupo homogéneo había cuatro ovejas (según el peso) y cada una recibió una de las cuatro dietas en orden aleatorio. Se midió el porcentaje de materia seca digerida de las 24 ovejas. Los siguientes son los datos:

	Grupo por peso (bloque)					
Dieta	**1**	**2**	**3**	**4**	**5**	**6**
Control	32.68	36.22	36.36	40.95	34.99	33.89
Urea en la alimentación	35.90	38.73	37.55	34.64	37.36	34.35
Tratada con amoniaco	49.43	53.50	52.86	45.00	47.20	49.76
Tratada con urea	46.58	42.82	45.41	45.08	43.81	47.40

a) Use un análisis de bloques completos aleatorizados para probar las diferencias entre las dietas. Emplee $\alpha = 0.05$.

b) Utilice la prueba de Dunnett para comparar las tres dietas con el control. Utilice $\alpha = 0.05$.

c) Dibuje una gráfica de probabilidad normal de los residuales.

13.47 En un estudio realizado por el personal del Departamento de Bioquímica de Virginia Tech se dieron tres dietas a un grupo de ratas con el objetivo de estudiar el efecto de cada una sobre el zinc dietético residual en el torrente sanguíneo. Se asignaron al azar cinco ratas preñadas a cada grupo dietético, y cada una recibió la dieta en el día 22 del embarazo. Se midió la cantidad de zinc en partes por millón. Los datos son los que siguen:

Dieta	1	0.50	0.42	0.65	0.47	0.44
	2	0.42	0.40	0.73	0.47	0.69
	3	1.06	0.82	0.72	0.72	0.82

Determine si hay una diferencia significativa en el zinc dietético residual entre las tres dietas. Use $\alpha = 0.05$. Lleve a cabo un ANOVA de un solo factor.

13.48 Se realizó un experimento para comparar tres tipos de pintura para buscar evidencia de diferencias en su calidad de desgaste. Las pinturas se expusieron a acciones abrasivas y se registró el tiempo, en horas, que tardaba en observarse la abrasión. Se usaron seis especímenes para cada tipo de pintura. Los datos son los siguientes:

Tipo de pintura									
1			2			3			
158	97	282	515	264	544	317	662	213
315	220	115	525	330	525	536	175	614

a) Realice un análisis de varianza para determinar si la evidencia sugiere que la calidad del desgaste de las tres pinturas es diferente. Utilice un valor *P* en sus conclusiones.

b) Si se encuentran diferencias significativas, diga cuáles son. ¿Hay alguna pintura que destaque? Analice sus hallazgos.

c) Haga todos los análisis gráficos que necesite para determinar si son válidas las suposiciones que se hicieron en el inciso *a*. Analice sus hallazgos.

d) Suponga que se determina que los datos para cada tratamiento tienen una distribución exponencial. ¿Sugiere esto un análisis alternativo? Si fuera así, hágalo y presente sus hallazgos.

13.49 Una empresa que troquela juntas de hojas de caucho, plástico y corcho desea comparar el número medio de juntas producidas por hora para los tres tipos de material. Se eligieron al azar dos máquinas troqueladoras como bloques. Los datos representan el número de juntas (en miles) producidas por hora. En la figura 13.16 de la página 557 se observa la salida de resultados del análisis.

Máquina	Material: Corcho			Caucho			Plástico		
A	4.31	4.27	4.40	3.36	3.42	3.48	4.01	3.94	3.89
B	3.94	3.81	3.99	3.91	3.80	3.85	3.48	3.53	3.42

a) ¿Por qué se eligieron las máquinas troqueladoras como bloques?

b) Grafique las seis medias para las combinaciones de máquinas y materiales.

c) ¿Hay un material que sea mejor?

d) ¿Existe interacción entre los tratamientos y los bloques? Si es así, diga si la interacción ocasiona alguna dificultad seria para llegar a una conclusión adecuada. Explique su respuesta.

13.50 Se hizo un estudio para comparar el rendimiento de tres marcas de gasolina competidoras. Se seleccionaron al azar cuatro modelos de automóvil de tamaño variable. A continuación se presentan los datos, en millas por galón. El orden de prueba es aleatorio para cada modelo.

Modelo	Marca de gasolina: *A*	*B*	*C*
A	32.4	35.6	38.7
B	28.8	28.6	29.9
C	36.5	37.6	39.1
D	34.4	36.2	37.9

a) Analice la necesidad de utilizar más de un solo modelo de automóvil.

b) Considere el ANOVA de la salida de resultados del *SAS* en la figura 13.17. ¿Es importante la marca de la gasolina?

c) ¿Qué marca de gasolina seleccionaría usted? Consulte el resultado de la prueba de Duncan.

13.51 Se utilizaron cuatro localidades diferentes del noreste para hacer mediciones de ozono, en partes por millón. Se recolectaron las cantidades de ozono en cinco muestras de cada localidad.

Localidad			
1	2	3	4
0.09	0.15	0.10	0.10
0.10	0.12	0.13	0.07
0.08	0.17	0.08	0.05
0.08	0.18	0.08	0.08
0.11	0.14	0.09	0.09

a) ¿Hay información suficiente que sugiera que existen diferencias en los niveles medios de ozono entre las diferentes localidades? Guíese usando un valor *P*.

b) Si se encuentran diferencias significativas en el inciso *a*, determine su naturaleza. Emplee cualesquiera métodos que haya aprendido.

13.52 Demuestre que el cuadrado medio del error

$$s^2 = \frac{SCE}{k(n-1)}$$

para el análisis de varianza en la clasificación de un factor es un estimado no sesgado de σ^2.

13.53 Demuestre el teorema 13.2.

13.54 Demuestre que la fórmula para calcular la *SCB*, en el análisis de varianza del diseño de bloques completos aleatorizados, es equivalente al término correspondiente en la identidad del teorema 13.3.

13.55 Para el diseño de bloques aleatorizados con k tratamientos y b bloques, demuestre que

$$E(SCB) = (b-1)\sigma^2 + k\sum_{j=1}^{b}\beta_j^2.$$

```
                          The GLM Procedure

Dependent Variable: gasket
                              Sum of
Source                DF      Squares    Mean Square   F Value    Pr > F
Model                  5   1.68122778     0.33624556     76.52    <.0001
Error                 12   0.05273333     0.00439444
Corrected Total       17   1.73396111

R-Square      Coeff Var       Root MSE     gasket Mean
0.969588       1.734095       0.066291         3.822778

Source                DF  Type III SS     Mean Square   F Value     Pr > F
material               2   0.81194444      0.40597222     92.38     <.0001
machine                1   0.10125000      0.10125000     23.04     0.0004
material*machine       2   0.76803333      0.38401667     87.39     <.0001

Level of      Level of             ------------gasket-----------
material      machine     N               Mean           Std Dev
cork          A           3          4.32666667        0.06658328
cork          B           3          3.91333333        0.09291573
plastic       A           3          3.94666667        0.06027714
plastic       B           3          3.47666667        0.05507571
rubber        A           3          3.42000000        0.06000000
rubber        B           3          3.85333333        0.05507571

Level of            ------------gasket-----------
material      N             Mean           Std Dev
cork          6       4.12000000        0.23765521
plastic       6       3.71166667        0.26255793
rubber        6       3.63666667        0.24287171

Level of            ------------gasket-----------
machine       N             Mean           Std Dev
A             9       3.89777778        0.39798800
B             9       3.74777778        0.21376259
```

Figura 13.16: Salida de resultados por computadora del *SAS* para el ejercicio de repaso 13.49.

```
                          The GLM Procedure
Dependent Variable: MPG
                             Sum of
 Source            DF       Squares   Mean Square   F Value    Pr > F
 Model              5   153.2508333    30.6501667     24.66    0.0006
 Error              6     7.4583333     1.2430556
 Corrected Total   11   160.7091667

 R-Square     Coeff Var      Root MSE      MPG Mean
 0.953591      3.218448      1.114924      34.64167

 Source            DF   Type III SS   Mean Square  F Value    Pr > F
 Model              3   130.3491667    43.4497222    34.95    0.0003
 Brand              2    22.9016667    11.4508333     9.21    0.0148
                Duncan's Multiple Range Test for MPG
 NOTE: This test controls the Type I comparisonwise error rate, not
 the experimentwise error rate.

                Alpha                        0.05
                Error Degrees of Freedom        6
                Error Mean Square        1.243056

                Number of Means          2         3
                Critical Range       1.929     1.999
      Means with the same letter are not significantly different.
            Duncan Grouping          Mean      N    Brand
                          A       36.4000      4    C
                          A
                       B  A       34.5000      4    B
                       B
                       B           33.0250      4    A
```

Figura 13.17 Salida de resultados por computadora del *SAS* para el ejercicio de repaso 13.50.

13.56 Proyecto de grupo: Resulta de interés determinar qué tipo de pelota deportiva se puede lanzar a la mayor distancia. La competencia incluye una pelota de tenis, una de beisbol y una de softbol. Divida el grupo en equipos de cinco estudiantes. Cada equipo debe diseñar y realizar un experimento separado, también debe analizar los datos de su propio experimento. Los cinco miembros del equipo lanzarán cada pelota (después de calentar el brazo el tiempo adecuado). La respuesta experimental será la distancia (en pies) que se lanza la pelota. Los datos de cada equipo incluirán 15 observaciones. Aspectos importantes:

a) No se trata de una competencia entre equipos. La competencia es entre los tres tipos de pelotas. Se esperaría que las conclusiones de cada equipo sean similares.

b) En cada equipo debe haber hombres y mujeres.

c) El diseño experimental de cada equipo deberá ser un diseño de bloques completos aleatorizados. Los cinco individuos que lanzan la pelota son los bloques.

d) Asegúrese de incorporar la aleatorización adecuada para realizar el experimento.

e) Los resultados deberán contener una descripción del experimento con una tabla de ANOVA que incluya un valor *P* y las conclusiones apropiadas. Se usarán técnicas gráficas y comparaciones múltiples en caso de ser necesarias. Saquen conclusiones prácticas con respecto a las diferencias entre los tipos de pelotas. Sean meticulosos.

13.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Al igual que otros procedimientos estudiados en capítulos anteriores, el análisis de varianza es razonablemente robusto con respecto a la suposición de normalidad, pero no lo es tanto en cuanto a la suposición de varianza homogénea. También observamos que la prueba de Bartlett para varianzas iguales es sumamente débil en relación con la normalidad.

Este capítulo es sumamente importante, ya que se trata de un punto "de inicio" para temas importantes, como el diseño de experimentos y el análisis de varianza. En el capítulo 14 se tratan los mismos temas, pero en los casos de extensiones a más de un factor y el análisis más complicado por la interpretación de la interacción entre factores. Hay ocasiones en que el papel de la interacción en un experimento científico es más importante que el papel de los factores principales (efectos principales). Ante la presencia de interacciones se hace un énfasis aún mayor en las técnicas gráficas. En los capítulos 14 y 15 será necesario proporcionar más detalles acerca del proceso de aleatorización, ya que el número de combinaciones de factores puede ser muy grande.

Capítulo 14

Experimentos factoriales (dos o más factores)

14.1 Introducción

Considere una situación en la que haya interés por estudiar el efecto de **dos factores**, *A* y *B*, sobre alguna respuesta. Por ejemplo, en un experimento químico nos gustaría variar en forma simultánea la presión de reacción y el tiempo de reacción, y estudiar el efecto que cada uno tiene sobre el producto. En un experimento biológico resulta de interés estudiar el efecto que tienen el tiempo de secado y la temperatura sobre la cantidad de sólidos (porcentaje por peso) que queda en muestras de levadura. Igual que en el capítulo 13, el término **factor** se utiliza en un sentido general para denotar cualquier característica del experimento que pueda variar de un ensayo a otro, como la temperatura, el tiempo o la presión. Los **niveles** de un factor se definen como los valores reales que se utilizan en el experimento.

Para cada uno de estos casos es importante determinar no sólo si cada uno de los 2 factores influye en la respuesta, sino también si hay una interacción significativa entre ellos. En lo que se refiere a la terminología, el experimento descrito aquí es de 2 factores, y el diseño experimental podría ser uno completamente aleatorizado, en el que las distintas combinaciones de tratamiento se asignan al azar a todas las unidades experimentales, o bien, un diseño de bloques completos aleatorizados, donde las combinaciones de factores se asignan al azar dentro de los bloques. En el ejemplo de la levadura, si se empleara un diseño completamente aleatorizado, las distintas combinaciones de tratamientos de temperatura y tiempo de secado se asignarían al azar a las muestras de levadura.

Muchos de los conceptos que se estudiaron en el capítulo 13 se extienden a 2 y 3 factores en este capítulo. El objetivo principal de este material es el uso del diseño completamente aleatorizado con un *experimento factorial*. Un experimento factorial con 2 factores implica ensayos experimentales (o uno solo) con todas las combinaciones de factores. Por ejemplo, en el caso de la temperatura y tiempo de secado con, digamos, 3 niveles de cada uno y $n = 2$ corridas por cada una de las 9 combinaciones, tendríamos un *experimento factorial de 2 factores en un diseño completamente aleatorizado*. Ninguno de ellos es un factor de bloqueo; nos interesa la manera en que cada uno influye en el porcentaje de sólidos en las muestras, y si interactúan o no. El biólogo dispondría de 18 muestras físicas de material que constituyen unidades experimentales. Luego, éstas se asignarían al azar a las 18 combinaciones (9 combinaciones de tratamiento, cada una de ellas por duplicado).

Antes de entrar en detalles analíticos, sumas de cuadrados y demás, sería interesante que el lector observe la clara conexión que existe entre lo que hemos descrito y la situación con el problema de un solo factor. Considere el experimento de la levadura. La explicación de los grados de libertad ayuda a que el lector o el analista visualicen la extensión. En un inicio, las 9 combinaciones de tratamientos deberían considerarse como si representaran un factor con 9 niveles (8 grados de libertad). Así, un vistazo inicial a los grados de libertad arroja lo siguiente:

Combinaciones de tratamiento	8
Error	9
Total	17

Efectos principales e interacción

En realidad el experimento se podría analizar como se describe en la tabla anterior. Sin embargo, es probable que la prueba F para las combinaciones no dé al analista la información que desea, es decir, el papel de la temperatura y del tiempo de secado. Tres tiempos de secado tienen asociados 2 grados de libertad, y a 3 temperaturas se asocian también 2 grados de libertad. Los factores principales, la temperatura y el tiempo de secado reciben el nombre de **efectos principales**, los cuales representan 4 de los 8 grados de libertad para las *combinaciones de factores*. Los 4 grados de libertad adicionales se asocian con la *interacción* entre los 2 factores. Como resultado, el análisis incluye

Combinaciones		8
Temperatura	2	
Tiempo de secado	2	
Interacción	4	
Error		9
Total		17

En el capítulo 13 vimos que en un análisis de varianza los factores pueden considerarse fijos o aleatorios, dependiendo del tipo de inferencia deseada y de la manera en que se eligieron los niveles. Aquí debemos considerar los efectos fijos, los efectos aleatorios e incluso los casos en que los efectos son mixtos. Conforme avancemos en estos temas pondremos mayor atención a los cuadrados medios esperados. En la siguiente sección nos centraremos en el concepto de interacción.

14.2 Interacción en el experimento de dos factores

En el modelo de bloques aleatorizados que se estudió previamente se supuso que en cada bloque se toma una observación de cada tratamiento. Si la suposición del modelo es correcta, es decir, si los bloques y los tratamientos son los únicos efectos reales y no hay interacción, el valor esperado del cuadrado medio del error es la varianza del error experimental σ^2. Sin embargo, suponga que existe interacción entre los tratamientos y los bloques, como lo indica el modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

de la sección 13.8. El valor esperado del cuadrado medio del error entonces es dado por

$$E\left[\frac{SCE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2 .$$

Los efectos del tratamiento y los bloques no aparecen en el cuadrado medio del error esperado, pero los efectos de la interacción sí. Entonces, si en el modelo hay interacción, el cuadrado medio del error refleja variación debida al error experimental más una contribución de la interacción y, para este plan experimental, no hay forma de separarlos.

La interacción y la interpretación de los efectos principales

Desde el punto de vista del experimentador, parecería necesario llegar a una prueba significativa sobre la existencia de una interacción, al separar la variación del error verdadero de aquel que se debe a la interacción. Los efectos principales, *A* y *B*, adoptan un significado distinto en presencia de la interacción. En el ejemplo biológico anterior el efecto que tiene el tiempo de secado sobre la cantidad de sólidos que quedan en la levadura muy bien podría depender de la temperatura a la que se expusieron las muestras. En general, podrían existir situaciones experimentales en las que el factor *A* tuviera un efecto positivo sobre la respuesta en un nivel del factor *B*; en tanto que con un nivel distinto de *B* el efecto de *A* sería negativo. Aquí se usa el término **efecto positivo** para indicar que el producto o la respuesta se incrementan conforme los niveles de un factor dado aumentan de acuerdo con cierto orden definido. En el mismo sentido, un **efecto negativo** corresponde a una disminución de la respuesta al aumentar los niveles del factor.

Considere, por ejemplo, los siguientes datos de temperatura (factor *A* con niveles t_1, t_2 y t_3 (en orden creciente) y tiempo de secado d_1, d_2 y d_3 (también en orden creciente). La respuesta es el porcentaje de sólidos. Estos datos son completamente hipotéticos y se dan para ilustrar un aspecto.

	B			
A	d_1	d_2	d_3	**Total**
t_1	4.4	8.8	5.2	18.4
t_2	7.5	8.5	2.4	18.4
t_3	9.7	7.9	0.8	18.4
Total	21.6	25.2	8.4	55.2

Es evidente que el efecto de la temperatura sobre el porcentaje de sólidos es positivo para el tiempo breve de secado d_1, pero negativo para el tiempo prolongado d_3. Esta **interacción clara** entre la temperatura y el tiempo de secado es evidentemente interesante para el biólogo; sin embargo, con base en los totales de las respuestas para las temperaturas t_1, t_2 y t_3, la suma de cuadrados de la temperatura, *SCT*, producirá un valor de 0. Entonces, se dice que la presencia de la interacción **enmascara** el efecto de la temperatura. Por ello, si se considera el efecto medio de la temperatura, promediado para el tiempo de secado, **no existe efecto alguno**. Entonces, esto define el efecto principal. Pero, por supuesto, es probable que esto no sea pertinente para el biólogo.

Antes de sacar cualquier conclusión final a partir de las pruebas de significancia sobre los efectos principales y los efectos de la interacción, el **experimentador debería observar primero si la prueba para la interacción es o no significativa**. Si la

interacción no es significativa, entonces los resultados de las pruebas sobre los efectos principales son importantes. No obstante, si la interacción debe ser significativa, entonces solamente son importantes aquellas pruebas sobre los efectos principales que resultan significativas. En presencia de una interacción, los efectos principales no significativos bien podrían ser resultado de enmascaramiento e indicar la necesidad de observar la influencia de cada factor a niveles fijos del otro.

Representación gráfica de la interacción

La presencia de interacción, así como su impacto científico, se puede interpretar adecuadamente usando **gráficas de interacción**. Las gráficas proporcionan una clara imagen de la tendencia de los datos para mostrar el efecto que tiene el cambio de un factor conforme se pasa de un nivel a otro del segundo factor. La figura 14.1 ilustra la fuerte interacción entre la temperatura y el tiempo de secado. La interacción se revela en las líneas no paralelas.

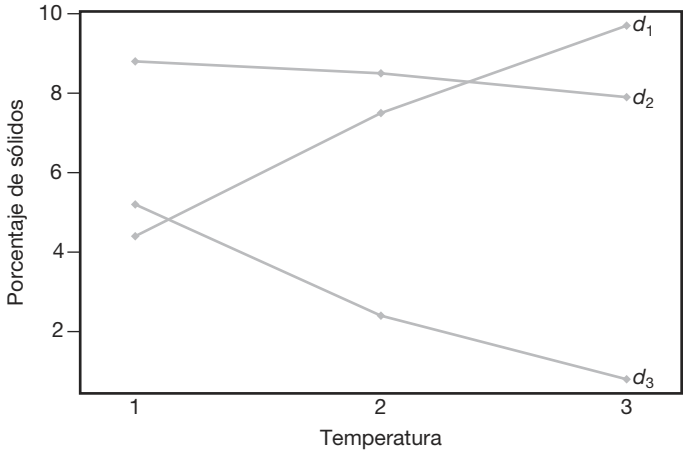


Figura 14.1: Gráfica de la interacción para los datos de temperatura y de tiempo de secado.

El *efecto relativamente fuerte de la temperatura* sobre el porcentaje de sólidos en el tiempo de secado más breve se refleja en la marcada pendiente de d_1. En el tiempo de secado medio, d_2, la temperatura tiene muy poco efecto, mientras que en el tiempo de secado prolongado d_3 la pendiente negativa indica un efecto negativo de la temperatura. Las gráficas de interacción como ésta le permiten al científico hacer una interpretación rápida y significativa de la interacción que existe. Debe ser evidente que el **paralelismo** en las gráficas indica la **ausencia de interacción**.

Necesidad de observaciones múltiples

En el experimento de 2 factores, la interacción y el error experimental sólo se separan si se hacen observaciones múltiples con las distintas combinaciones de tratamiento. Para máxima eficiencia debe haber el mismo número n de observaciones para cada combinación. Éstas deben ser verdaderas réplicas, no sólo medidas repetidas. Por ejemplo, en

el caso de la levadura, si para cada combinación de temperatura y tiempo de secado se toman $n = 2$ observaciones, debería haber dos muestras separadas y no sólo mediciones repetidas en la misma muestra. Esto permite que la variabilidad debida a las unidades experimentales aparezca en el "error", de manera que la variación no es un simple error de medición.

14.3 Análisis de varianza de dos factores

Al presentar las fórmulas generales para el análisis de varianza de un experimento de 2 factores utilizando observaciones repetidas en un diseño completamente aleatorizado, debe considerarse el caso de n réplicas de las combinaciones del tratamiento, determinadas por a niveles del factor A y b niveles del factor B. Las observaciones se podrían clasificar usando un arreglo rectangular, donde los renglones representan los niveles del factor A y las columnas representan los niveles del factor B. Cada combinación de tratamiento define una celda del arreglo. Así, se tienen ab celdas, cada una de las cuales contiene n observaciones. Se denota con y_{ijk}, la k-ésima observación tomada en el i-ésimo nivel del factor A y el j-ésimo nivel del factor B. En la tabla 14.1 se muestran las abn observaciones.

Tabla 14.1: Experimento de dos factores con n réplicas

	B					
A	**1**	**2**	$\cdots$	***b***	**Total**	**Media**
1	y_{111}	y_{121}	$\cdots$	y_{1b1}	$Y_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	$\cdots$	y_{1b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{11n}	y_{12n}	$\cdots$	y_{1bn}		
2	y_{211}	y_{221}	$\cdots$	y_{2b1}	$Y_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	$\cdots$	y_{2b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}		
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}	$Y_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}		
Total	$Y_{.1.}$	$Y_{.2.}$	$\cdots$	$Y_{.b.}$	$Y_{...}$	
Media	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$		$\bar{y}_{...}$

Las observaciones en la celda (ij)-ésima constituyen una muestra aleatoria de tamaño n de una población que se supone tiene distribución normal con media μ_{ij} y varianza σ^2. Se supone que todas las ab poblaciones tienen la misma varianza σ^2. Se

definen los siguientes símbolos útiles, algunos de los cuales se utilizan en la tabla 14.1:

$Y_{ij.}$ = suma de las observaciones en la (ij)-ésima celda,
$Y_{i..}$ = suma de las observaciones para el i-ésimo nivel del factor A,
$Y_{.j.}$ = suma de las observaciones para el j-ésimo nivel del factor B,
$Y_{...}$ = suma de todas las abn observaciones,
$\bar{y}_{ij.}$ = media de las observaciones en la (ij)-ésima celda,
$\bar{y}_{i..}$ = media de las observaciones para el i-ésimo nivel del factor A,
$\bar{y}_{.j.}$ = media de las observaciones para el j-ésimo nivel del factor B,
$\bar{y}_{...}$ = media de todas las abn observaciones.

A diferencia de la situación para un solo factor, que se cubrió con amplitud en el capítulo 13, en éste supondremos que las **poblaciones**, de las que se toman n observaciones independientes con distribución idéntica, son **combinaciones** de los factores. Asimismo, se supondrá siempre que de cada combinación de factores se toma un número igual (n) de observaciones. En los casos en que los tamaños de las muestras por combinación son desiguales, los cálculos son más complicados, aunque los conceptos son transferibles.

Modelo e hipótesis para el problema de dos factores

Cada observación de la tabla 14.1 se puede escribir en la siguiente forma:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

donde ϵ_{ijk} mide las desviaciones de los valores y_{ijk} observados en la (ij)-ésima celda a partir de la media de la población μ_{ij}. Si $(\alpha\beta)_{ij}$ denota el efecto de la interacción del i-ésimo nivel del factor A y el j-ésimo nivel del factor B, α_i el efecto del i-ésimo nivel del factor A, β_j el efecto del j-ésimo nivel del factor B, y μ la media conjunta, escribimos

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

y, entonces,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

a las que se imponen las restricciones

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$$

Las 3 hipótesis por probar son las siguientes:

1. H_0': $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$,
H_1': Al menos una de las α_i no es igual a 0.

2. H_0'': $\beta_1 = \beta_2 = \cdots = \beta_b = 0$,
H_1'': Al menos una de las β_j no es igual a 0.

3. H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0$,
H_1''': Al menos una de las $(\alpha\beta)_{ij}$ no es igual a 0.

Se alerta al lector acerca del problema del enmascaramiento de los efectos principales cuando la interacción contribuye de manera importante en el modelo. Se recomienda considerar primero el resultado de la prueba de interacción y, luego, la interpretación de la prueba del efecto principal; la naturaleza de la conclusión científica depende de si se encontró interacción. Si ésta se descarta, entonces se pueden probar las hipótesis 1 y 2 y la interpretación es muy sencilla. Sin embargo, si se descubre que hay interacción, la interpretación puede ser más complicada, como se vio al analizar el tiempo de secado y la temperatura en la sección previa. La estructura de las pruebas de hipótesis 1, 2 y 3 se estudiará en las secciones siguientes. En el análisis del ejemplo 14.1 se incluirá la interpretación de los resultados.

Las pruebas de las hipótesis anteriores se basarán en la comparación de estimados independientes de σ^2, obtenidos al separar la suma de cuadrados total de los datos en 4 componentes mediante la siguiente identidad.

Partición de la variabilidad en el caso de dos factores

Teorema 14.1: Identidad de la suma de cuadrados

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\ldots})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{\ldots})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{\ldots})^2$$
$$+ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\ldots})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$$

Simbólicamente, la identidad de la suma de cuadrados se escribe como

$$SCT = SCA + SCB + SC(AB) + SCE$$

donde a *SCA* y *SCB* se les denomina la suma de cuadrados para los efectos principales *A* y *B*, respectivamente, *SC*(*AB*) recibe el nombre de suma de cuadrados de la interacción para *A* y *B*, y *SCE* es la suma de cuadrados del error. La partición de los grados de libertad se efectúa de acuerdo con la identidad

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1).$$

Formación de los cuadrados medios

Si dividimos cada una de las sumas de cuadrados en el lado derecho de la identidad de la suma de cuadrados entre su número correspondiente de grados de libertad, obtenemos los cuatro estadísticos

$$S_1^2 = \frac{SCA}{a-1}, \qquad S_2^2 = \frac{SCB}{b-1}, \qquad S_3^2 = \frac{SC(AB)}{(a-1)(b-1)}, \qquad S^2 = \frac{SCE}{ab(n-1)}.$$

Todos estos estimados de la varianza son estimados independientes de σ^2, siempre que no haya efectos α_i, β_j ni, por supuesto, $(\alpha\beta)_{ij}$. Si las sumas de cuadrados se interpretan

como funciones de las variables aleatorias independientes $y_{111}, y_{112}, \ldots, y_{abn}$, no es difícil comprobar que

$$E(S_1^2) = E\left[\frac{SCA}{a-1}\right] = \sigma^2 + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2,$$

$$E(S_2^2) = E\left[\frac{SCB}{b-1}\right] = \sigma^2 + \frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

$$E(S_3^2) = E\left[\frac{SC(AB)}{(a-1)(b-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2,$$

$$E(S^2) = E\left[\frac{SCE}{ab(n-1)}\right] = \sigma^2,$$

a partir de lo cual se observa de inmediato que los 4 estimados de σ^2 son no sesgados cuando H_0', H_0'', y H_0''' son verdaderas.

Para probar la hipótesis H_0', de que los efectos de los factores A son todos iguales a cero, se calcula la siguiente razón:

Prueba F para el factor A

$$f_1 = \frac{s_1^2}{s^2},$$

que es un valor de la variable aleatoria F_1, el cual tiene la distribución F con $a-1$ y $ab(n-1)$ grados de libertad cuando H_0', es verdadera. La hipótesis nula se rechaza al nivel de significancia α cuando $f_1 > f_\alpha[a-1, ab(n-1)]$.

De manera similar, para probar la hipótesis H_0'', de que todos los efectos del factor B son iguales a cero, se calcula la razón:

Prueba F para el factor B

$$f_2 = \frac{s_2^2}{s^2},$$

que es un valor de la variable aleatoria F_2 que tiene la distribución F con $b-1$ y $ab(n-1)$ grados de libertad cuando H_0'', es verdadera. Esta hipótesis se rechaza al nivel de significancia α cuando $f_2 > f_\alpha[b-1, ab(n-1)]$.

Por último, para probar la hipótesis H_0''', de que todos los efectos de interacción son iguales a 0, se calcula la razón siguiente:

Prueba F para la interacción

$$f_3 = \frac{s_3^2}{s^2},$$

que es un valor de la variable aleatoria F_3, el cual tiene la distribución F con $(a-1)(b-1)$ y $ab(n-1)$ grados de libertad cuando H_0''' es verdadera. Concluimos que, a un nivel de significancia α, hay interacción cuando $f_\alpha[(a-1)(b-1), ab(n-1)]$.

Como se indicó en la sección 14.2, se recomienda interpretar la prueba para la interacción antes de tratar de hacer inferencias sobre los efectos principales. Si la interacción no es significativa, entonces hay evidencia de que las pruebas sobre los efectos principales son interpretables. El rechazo de la hipótesis 1 de la página 566 implica que las medias de la respuesta en los niveles del factor A difieren significativamente, mientras que

el rechazo de la hipótesis 2 implica una condición similar para las medias en los niveles del factor B. Sin embargo, una interacción significativa podría muy bien implicar que los datos se deberían analizar de una manera un poco diferente, **quizá observando el efecto del factor A en niveles fijos del factor B**, y así sucesivamente.

Los cálculos en un problema de análisis de varianza para un experimento de 2 factores con n réplicas suelen resumirse como se ilustra en la tabla 14.2.

Tabla 14.2: Análisis de varianza para el experimento de 2 factores con n réplicas

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Efecto principal				
A	SCA	$a-1$	$s_1^2 = \frac{SCA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SCB	$b-1$	$s_2^2 = \frac{SCB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Interacciones de 2 factores				
AB	$SC(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SC(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SCE	$ab(n-1)$	$s^2 = \frac{SCE}{ab(n-1)}$	
Total	STC	$abn-1$		

Ejemplo 14.1: En un experimento realizado para determinar cuál de 3 sistemas de misiles distintos es preferible, se midió la tasa de combustión del propulsor para 24 arranques estáticos. Se emplearon 4 tipos de combustible diferentes y el experimento generó observaciones duplicadas de las tasas de combustión para cada combinación de los tratamientos.

Los datos, ya codificados, se presentan en la tabla 14.3. Pruebe las siguientes hipótesis: *a*) H_0': no hay diferencia en las tasas medias de combustión del propulsor cuando se emplean diferentes sistemas de misiles, *b*) H_0'': no existe diferencia en las tasas medias de combustión de los 4 tipos de propulsor, *c*) H_0''': no hay interacción entre los distintos sistemas de misiles y los diferentes tipos de propulsor.

Tabla 14.3: Tasas de combustión del propulsor

Sistema de misiles	Tipo de propulsor			
	b_1	b_2	b_3	b_4
a_1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
a_2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
a_3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1

***Solución*:** **1.** *a*) H_0': $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

b) H_0'': $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

c) H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{34} = 0$.

2. *a*) H_1': Al menos una de las α_i no es igual a 0.
b) H_1'': Al menos una de las β_j no es igual a 0.
c) H_1''': Al menos una de las $(\alpha\beta)_{ij}$ no es igual a 0.

Se utiliza la fórmula de la suma de cuadrados que se describió en el teorema 14.1. En la tabla 14.4 se presenta el análisis de varianza.

Tabla 14.4: Análisis de varianza para los datos de la tabla 14.3

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**
Sistema de misiles	14.52	2	7.26	5.84
Tipo de propulsor	40.08	3	13.36	10.75
Interacción	22.16	6	3.69	2.97
Error	14.91	12	1.24	
Total	91.68	23		

Se remite al lector al procedimiento de modelos lineales generales (GLM, por sus siglas en inglés) del *SAS* para el análisis de los datos de la tasa de combustión de la figura 14.2. Observe la forma en que al principio se prueba el "modelo" (11 grados de libertad), y por separado se prueban el sistema, el tipo y el sistema por tipo de interacción. La prueba *f* en el modelo ($P = 0.0030$) prueba la acumulación de los 2 efectos principales y la interacción.

a) Rechace H_0' y concluya que los distintos sistemas de misiles resultan en diferentes tasas medias de combustión del propulsor. El valor *P* es de aproximadamente 0.0169.

b) Rechace H_0'' y concluya que las tasas medias de combustión del propulsor no son las mismas para los 4 tipos de propulsores. El valor *P* es de aproximadamente 0.0010.

c) La interacción es apenas insignificante al nivel 0.05, pero el valor *P* de aproximadamente 0.0513 indicaría que la interacción debe tomarse en serio.

En este momento debemos hacer algún tipo de interpretación de la interacción. Debe destacarse que la significancia estadística de un efecto principal tan sólo implica que las *medias marginales son significativamente diferentes*. Sin embargo, considere la tabla de promedios de 2 factores de la tabla 14.5.

Tabla 14.5: Interpretación de la interacción

	b_1	b_2	b_3	b_4	**Promedio**
a_1	33.35	31.45	28.25	28.95	30.50
a_2	32.60	30.00	28.40	27.70	29.68
a_3	28.85	28.10	28.50	28.95	28.60
Promedio	31.60	29.85	28.38	28.53	

Es evidente que hay más información importante en el cuerpo de la tabla, tendencias que son inconsistentes con la tendencia que describe los promedios marginales. La tabla 14.5 sugiere con certeza que el efecto del tipo de propulsor depende del sistema que se utiliza. Por ejemplo, para el sistema 3, el efecto del tipo de propulsor no parece ser

```
                              The  GLM  Procedure
Dependent  Variable:  rate
                                 Sum of
 Source                DF       Squares     Mean Square     F Value    Pr > F
 Model                 11    76.76833333      6.97893939        5.62    0.0030
 Error                 12    14.91000000      1.24250000
 Corrected  Total      23    91.67833333

 R-Square       Coeff  Var       Root MSE       rate Mean
 0.837366        3.766854        1.114675        29.59167

 Source                DF    Type III SS     Mean Square    F Value     Pr > F
 system                 2    14.52333333      7.26166667       5.84     0.0169
 type                   3    40.08166667     13.36055556      10.75     0.0010
 system*type            6    22.16333333      3.69388889       2.97     0.0512
```

Figura 14.2: Salida de resultados del *SAS* para el análisis de los datos de la tasa de combustión del propulsor de la tabla 14.3.

importante, aunque tiene un efecto grande si se utiliza el sistema 1 o el 2. Esto explica la interacción "significativa" entre esos 2 factores. Más adelante se revelará más información acerca de esta interacción. ∎

Ejemplo 14.2: Remítase al ejemplo 14.1 y elija 2 contrastes ortogonales para dividir la suma de cuadrados del sistema de misiles en componentes con un solo grado de libertad, los cuales utilizará para comparar los sistemas 1 y 2 con el 3, y el sistema 1 contra el sistema 2.

***Solución*:** El contraste para comparar los sistemas 1 y 2 con el 3 es

$$w_1 = \mu_{1.} + \mu_{2.} - 2\mu_{3.}.$$

Un segundo contraste, ortogonal a w_1, para comparar el sistema 1 con el 2, es dado por $w_2 = \mu_{1.} - \mu_{2.}$. Las sumas de cuadrados con un solo grado de libertad son

$$SCw_1 = \frac{[244.0 + 237.4 - (2)(228.8)]^2}{(8)[(1)^2 + (1)^2 + (-2)^2]} = 11.80$$

y

$$SCw_2 = \frac{(244.0 - 237.4)^2}{(8)[(1)^2 + (-1)^2]} = 2.72.$$

Observe que $SCw_1 + SCw_2 = SCA$, como se esperaba. Los valores f calculados correspondientes a w_1 y w_2 son, respectivamente,

$$f_1 = \frac{11.80}{1.24} = 9.5 \qquad \text{y} \qquad f_2 = \frac{2.72}{1.24} = 2.2.$$

Al comparar con el valor crítico $f_{0.05}(1, 12) = 4.75$, se encuentra que f_1 es significativo. De hecho, el valor P es menor que 0.01. Así, el primer contraste indica que se rechaza la hipótesis

$$H_0\colon \frac{1}{2}(\mu_{1\cdot} + \mu_{2\cdot}) = \mu_{3\cdot}$$

Como $f_2 < 4.75$, las tasas medias de combustión del primer y segundo sistemas no son significativamente diferentes. ■

Impacto de la interacción significativa en el ejemplo 14.1

Si la hipótesis de que no hay interacción en el ejemplo 14.1 es verdadera, podríamos hacer las comparaciones *generales* del ejemplo 14.2 relacionado con los sistemas de misiles, en lugar de comparaciones separadas para cada propulsor. De manera similar, se podrían realizar comparaciones generales entre los propulsores, en vez de comparar por separado cada sistema de misiles. Por ejemplo, se podrían comparar los propulsores 1 y 2 con el 3 y 4, y también el 1 contra el 2. Las razones *f* resultantes, cada una con 1 y 12 grados de libertad, resultan ser de 24.81 y 7.39, respectivamente, y ambas son muy significativas al nivel 0.05.

Por los promedios de los propulsores, parece haber evidencia de que el 1 ofrece la tasa media de combustión más alta. Un experimentador prudente sería cauteloso al sacar conclusiones generales en un problema como éste, donde la razón *f* de la interacción está apenas por debajo del valor crítico de 0.05. Por ejemplo, la evidencia general, 31.60 contra 29.85 sobre el promedio para los 2 propulsores, indica con claridad que el 1 es superior al 2, en términos de una mayor tasa de combustión. Sin embargo, si nos restringimos al sistema 3, donde tenemos un promedio de 28.85 para el propulsor 1 en oposición a 28.10 para el propulsor 2, parece haber una diferencia mínima o incluso ninguna entre estos 2 propulsores. De hecho, parece que hay una estabilización de las tasas de combustión para los distintos propulsores si se opera con el sistema 3. Es claro que existe evidencia general que indica que el sistema 1 ofrece una tasa de combustión más alta que el sistema 3, pero parece que esta conclusión no se sostiene si nos restringimos al propulsor 4.

Para recabar evidencias concluyentes de que la interacción está *produciendo dificultades considerables en la obtención de conclusiones generales sobre los efectos principales*, el analista puede hacer una prueba *t* sencilla utilizando las tasas de combustión promedio del sistema 3. Considere una comparación del propulsor 1 contra el 2 usando únicamente el sistema 3. Se toma prestado un estimado de σ^2 del análisis general, es decir, se utiliza $s^2 = 1.24$ con 12 grados de libertad, y se obtiene

$$|t| = \frac{0.75}{\sqrt{2s^2/n}} = \frac{0.75}{\sqrt{1.24}} = 0.67,$$

que no se acerca a ser significativa. Esta ilustración sugiere que, en presencia de interacción, debería tenerse cautela con la interpretación estricta de los efectos principales.

Análisis gráfico para el problema de dos factores del ejemplo 14.1

Muchos de los mismos tipos de ilustraciones gráficas que se sugirió emplear en los problemas de un factor también se aplican en el caso de 2 factores. Las gráficas en 2 dimensiones de las medias de las celdas o de las medias de las combinaciones de tratamientos ofrecen información sobre la presencia de interacciones entre los 2 factores.

Además, una gráfica de los residuales contra los valores ajustados bien podría indicar si se cumple o no la suposición de la varianza homogénea. Por supuesto, es frecuente que una violación de la suposición de varianza homogénea implique un aumento en la varianza del error conforme *los números de la respuesta se vuelven más grandes*. Como resultado, esta gráfica podría resaltar la violación.

La figura 14.3 presenta la gráfica de las medias de las celdas para el caso del propulsor de los sistemas de misiles del ejemplo 14.1. Observe gráficamente (en este caso) cuánta falta de paralelismo hay. Note el aplanamiento de la parte de la figura que indica el efecto del propulsor para el sistema 3. Esto ilustra la interacción entre los factores. La figura 14.4 muestra la gráfica de los residuales contra los valores ajustados para los mismos datos. Al parecer no hay dificultades con la suposición de la varianza homogénea.

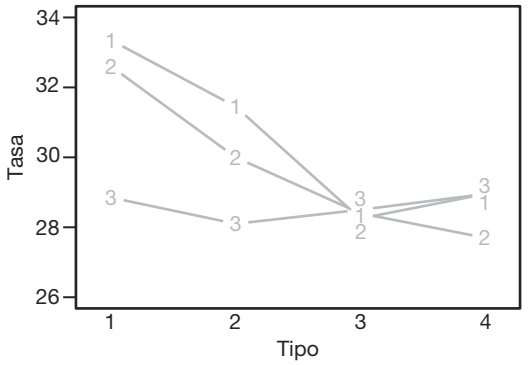


Figura 14.3: Gráfica de las medias de las celdas para los datos del ejemplo 14.1. Los números representan los sistemas de misiles.

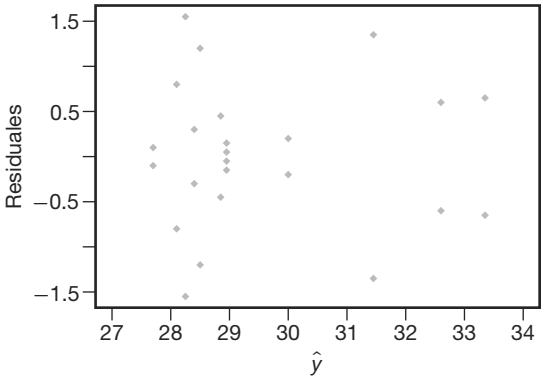


Figura 14.4: Gráfica de los residuales de los datos del ejemplo 14.1.

Ejemplo 14.3: Un ingeniero eléctrico investiga un proceso de grabado con plasma que se emplea en la fabricación de semiconductores. Es de interés estudiar los efectos de 2 factores, la cantidad de flujo (*A*) del gas C_2F_6 y la potencia aplicada al cátodo (*B*). La respuesta es la velocidad de grabado. Cada factor se aplica a 3 niveles y se hacen 2 corridas experimentales sobre la velocidad de grabado para cada una de las 9 combinaciones. El plan representa un diseño completamente aleatorizado. En la tabla 14.6 se presentan los datos. La velocidad de grabado se expresa en A°/min.

Tabla 14.6: Datos para el ejemplo 14.3

Cantidad de flujo del C_2F_6	**Potencia suministrada**		
	1	**2**	**3**
1	288	488	670
	360	465	720
2	385	482	692
	411	521	724
3	488	595	761
	462	612	801

Los niveles de los factores están en orden ascendente, donde el nivel 1 es el más bajo y el 3 el más alto.

a) Elabore una tabla de análisis de varianza y saque conclusiones; empiece con la prueba de interacción.

b) Haga pruebas sobre los efectos principales y saque conclusiones.

***Solución*:** En la figura 14.5 se muestra una salida de resultados por computadora del *SAS*. De ese listado se concluye lo siguiente.

```
                          The GLM Procedure
Dependent Variable: etchrate
                                 Sum of
Source                    DF       Squares     Mean Square    F Value    Pr > F
Model                      8    379508.7778      47438.5972      61.00    <.0001
Error                      9      6999.5000        777.7222
Corrected Total           17    386508.2778

R-Square       Coeff Var       Root MSE     etchrate Mean
0.981890        5.057714       27.88767           551.3889

Source                    DF    Type III SS     Mean Square    F Value    Pr > F
c2f6                       2     46343.1111      23171.5556      29.79    0.0001
power                      2    330003.4444     165001.7222     212.16    <.0001
c2f6*power                 4      3162.2222        790.5556       1.02    0.4485
```

Figura 14.5: Una salida de resultados por computadora del *SAS* para el ejemplo 14.3.

a) El valor *P* para la prueba de interacción es 0.4485. Se concluye que la interacción no es significativa.

b) Existe una diferencia significativa en la velocidad media de grabado para los 3 niveles de la velocidad de flujo del C_2F_6. Una prueba de Duncan muestra que la velocidad media de grabado para el nivel 3 es significativamente mayor que para el nivel 2, y la

velocidad para el nivel 2 es significativamente mayor que para el nivel 1. Véase la figura 14.6*a*.

Existe una diferencia significativa en la velocidad media de grabado basada en el nivel de potencia al cátodo. Una prueba de Duncan revela que la velocidad de grabado para el nivel 3 es significativamente más alta que para el 2, y que la velocidad para el nivel 2 es significativamente más alta que para el 1. Véase la figura 14.6*b*.

```
Duncan Grouping      Mean    N  c2f6
               A    619.83    6  3
               B    535.83    6  2
               C    498.50    6  1
                  (a)

Duncan Grouping      Mean    N  power
               A    728.00    6  3
               B    527.17    6  2
               C    399.00    6  1
                  (b)
```

Figura 14.6: Una salida de resultados por computadora del *SAS* para el ejemplo 14.3. *a*) Prueba de Duncan de la cantidad de flujo del gas; *b*) Prueba de Duncan de la potencia.

Ejercicios

14.1 Se realizó un experimento para estudiar los efectos de la temperatura y el tipo de horno sobre la vida de un componente en particular. En el experimento se utilizaron 4 tipos de horno y 3 niveles de temperatura. Se asignaron 24 piezas al azar, 2 para cada combinación de tratamientos y se registraron los siguientes resultados.

	Horno			
Temperatura (^oF)	O_1	O_2	O_3	O_4
500	227	214	225	260
	221	259	236	229
550	187	181	232	246
	208	179	198	273
600	174	198	178	206
	202	194	213	219

A un nivel de significancia de 0.05 pruebe las hipótesis de que

a) las diferentes temperaturas no tienen efecto en la vida del componente;
b) los diferentes hornos no tienen efecto en la vida del componente;
c) no hay interacción entre el tipo de horno y la temperatura.

14.2 El Departamento de Nutrición Humana y Alimentos del Virginia Tech realizó un estudio titulado *Vitamin C Retention in Reconstituted Frozen Orange Juice* sobre la estabilidad de la vitamina C en el concentrado de jugo de naranja congelado reconstituido y almacenado en un refrigerador durante un periodo de hasta una semana. Se probaron 3 tipos de concentrado de jugo de naranja congelado en 3 periodos distintos, los cuales se refieren al número de días transcurridos desde que se mezcló el jugo hasta que se probó. Se registraron los resultados en miligramos de ácido ascórbico por litro. Utilice un nivel de significancia de 0.05 para probar las hipótesis de que

a) no hay diferencias en el contenido de ácido ascórbico entre las diferentes marcas de concentrado de jugo de naranja;
b) no existen diferencias en el contenido de ácido ascórbico para distintos periodos;
c) no hay interacción entre las marcas de concentrado de jugo de naranja y el número de días transcurridos desde que el jugo se mezcló hasta que se probó.

	Tiempo (días)					
Marca	**0**		**3**		**7**	
Richfood	52.6	54.2	49.4	49.2	42.7	48.8
	49.8	46.5	42.8	53.2	40.4	47.6
Sealed-Sweet	56.0	48.0	48.8	44.0	49.2	44.0
	49.6	48.4	44.0	42.4	42.0	43.2
Minute Maid	52.5	52.0	48.0	47.0	48.5	43.3
	51.8	53.6	48.2	49.6	45.2	47.6

14.3 Se estudió el desempeño de 3 cepas de ratas en una prueba de laberintos en 2 condiciones ambientales diferentes. Se registraron las puntuaciones de error de las 48 ratas:

	Cepa					
Ambiente	**Brillante**		**Mezclada**		**Torpe**	
Libre	28	12	33	83	101	94
	22	23	36	14	33	56
	25	10	41	76	122	83
	36	86	22	58	35	23
Restringido	72	32	60	89	136	120
	48	93	35	126	38	153
	25	31	83	110	64	128
	91	19	99	118	87	140

Utilice un nivel de significancia de 0.01 para probar las hipótesis de que

a) no hay diferencia en las puntuaciones de error para ambientes diferentes;
b) no existe diferencia en las puntuaciones de error para cepas diferentes;
c) no hay interacción entre los ambientes y las cepas de las ratas.

14.4 La fatiga por corrosión de los metales se define como la acción simultánea de tensión cíclica y ataque químico sobre una estructura metálica. Una técnica muy utilizada para minimizar el daño de la fatiga por corrosión en el aluminio requiere la aplicación de un recubrimiento protector. En un estudio efectuado por el Departamento de Ingeniería Mecánica de Virginia Tech se utilizaron 3 niveles diferentes de humedad:

Bajo: 20 a 25% de humedad relativa

Medio: 55 a 60% de humedad relativa

Alto: 86 a 91% de humedad relativa

y 3 tipos de recubrimiento:

No revestido: Sin recubrimiento

Anodizado: Recubrimiento de óxido anódico por ácido sulfúrico

Conversión: Recubrimiento por conversión química de cromato.

Los datos de fatiga por corrosión, expresados en miles de ciclos hasta que se presenta la falla, se registraron como sigue:

	Humedad relativa					
Recubrimiento	**Baja**		**Media**		**Alta**	
	361	469	314	522	1344	1216
No revestido	466	937	244	739	1027	1097
	1069	1357	261	134	1011	1011
	114	1032	322	471	78	466
Anodizado	1236	92	306	130	387	107
	533	211	68	398	130	327
	130	1482	252	874	586	524
Conversión	841	529	105	755	402	751
	1595	754	847	573	846	529

a) Lleve a cabo un análisis de varianza con $\alpha = 0.05$ para probar si existen efectos principales y efectos de interacción.
b) Utilice la prueba de Duncan de rango múltiple a un nivel de significancia de 0.05 para determinar cuáles niveles de humedad relativa dan como resultado daños distintos de fatiga por corrosión.

14.5 Para determinar cuáles músculos necesitan someterse a un programa de acondicionamiento para mejorar el rendimiento individual en el servicio tendido que se usa en el tenis, el Departamento de Salud, Educación Física y Recreación de Virginia Tech realizó un estudio de 5 músculos diferentes:

1: deltoides anterior 4: deltoides medio
2: pectoral mayor 5: tríceps
3: deltoides posterior

los cuales se probaron en cada uno de 3 sujetos; el experimento se efectuó 3 veces para cada combinación de tratamiento. Los datos electromiográficos que se registraron durante el servicio se presentan a continuación.

	Músculo				
Sujeto	**1**	**2**	**3**	**4**	**5**
1	32	5	58	10	19
	59	1.5	61	10	20
	38	2	66	14	23
2	63	10	64	45	43
	60	9	78	61	61
	50	7	78	71	42
3	43	41	26	63	61
	54	43	29	46	85
	47	42	23	55	95

Utilice un nivel de significancia de 0.01 para probar las hipótesis de que

a) diferentes sujetos tienen medidas electromiográficas iguales;
b) los diferentes músculos no tienen un efecto en las medidas electromiográficas;
c) no hay interacción entre los sujetos y los tipos de músculos.

14.6 Se realizó un experimento para determinar si los aditivos incrementan la adherencia de productos de caucho. Se elaboraron 16 productos con el aditivo nuevo y otros 16 sin dicho aditivo. Se registró la siguiente adherencia.

	Temperatura (°C)			
	50	**60**	**70**	**80**
	2.3	3.4	3.8	3.9
Sin el aditivo	2.9	3.7	3.9	3.2
	3.1	3.6	4.1	3.0
	3.2	3.2	3.8	2.7
	4.3	3.8	3.9	3.5
Con el aditivo	3.9	3.8	4.0	3.6
	3.9	3.9	3.7	3.8
	4.2	3.5	3.6	3.9

Haga un análisis de varianza para probar la existencia de efectos principales y de interacción significativos.

14.7 Se sabe que la velocidad de extracción de cierto polímero depende de la temperatura de reacción y de la cantidad de catalizador empleada. Se hizo un experimento en 4 niveles de temperatura y 5 niveles de catalizador, y se registró la velocidad de extracción en la siguiente tabla:

	Cantidad de catalizador				
	0.5%	**0.6%**	**0.7%**	**0.8%**	**0.9%**
50°C	38	45	57	59	57
	41	47	59	61	58
60°C	44	56	70	73	61
	43	57	69	72	58
70°C	44	56	70	73	61
	47	60	67	61	59
80°C	49	62	70	62	53
	47	65	55	69	58

Realice un análisis de varianza. Pruebe si hay efectos principales y de interacción significativos.

14.8 En Myers, Montgomery y Anderson-Cook (2009) se estudia un escenario donde se describe un proceso de laminado por prensado. La respuesta es el espesor del material. Los factores que podrían afectar el espesor incluyen la cantidad de níquel (*A*) y el pH (*B*). Se diseñó un experimento con 2 factores. El plan es un diseño completamente aleatorizado en el que las prensas individuales se asignan al azar a las combinaciones de factores. En el experimento se utilizan 3 niveles de pH y 2 niveles de contenido de níquel. Los espesores, en cm $\times 10^{-3}$, son los siguientes:

Contenido de níquel (gramos)	**pH**		
	6	**5.5**	**6**
18	250	211	221
	195	172	150
	188	165	170
10	115	88	69
	165	112	101
	142	108	72

a) Elabore la tabla del análisis de varianza con pruebas para los efectos principales y de interacción. Incluya valores *P*.
b) Saque conclusiones para ingeniería. ¿Qué aprendió del análisis de estos datos?
c) Elabore una gráfica que ilustre la presencia o ausencia de interacción.

14.9 Un ingeniero está interesado en los efectos de la velocidad de corte y la geometría de la herramienta sobre las horas de vida de una máquina-herramienta. Se utilizan 2 velocidades de corte y 2 geometrías distintas. Se llevan a cabo 3 pruebas experimentales con cada una de las 4 combinaciones. Los datos son los siguientes:

Geometría de la herramienta	**Velocidad de corte**					
	Baja			**Alta**		
1	22	28	20	34	37	29
2	18	15	16	11	10	10

a) Calcule la tabla del análisis de varianza con pruebas sobre los efectos principales y de interacción.
b) Haga comentarios sobre el efecto que tiene la interacción sobre la prueba de la velocidad de corte.
c) Efectúe pruebas secundarias que permitan al ingeniero conocer el verdadero impacto de la velocidad de corte.
d) Construya una gráfica que ilustre el efecto de interacción.

14.10 En un experimento se estudiaron 2 factores de un proceso de manufactura de un circuito integrado. El propósito del experimento es conocer su efecto sobre la resistividad de las obleas de silicio. Los factores son la dosis del implante (2 niveles) y la posición de la caldera (3 niveles). El experimento es costoso, por lo que sólo se hizo una corrida con cada combinación. Los datos son los siguientes.

Dosis	**Posición**		
1	15.5	14.8	21.3
2	27.2	24.9	26.1

Se supone que no hay interacción entre esos 2 factores.
a) Escriba el modelo y explique sus términos.
b) Elabore la tabla de análisis de varianza.
c) Explique los 2 grados de libertad del "error".
d) Use una prueba de Tukey para hacer pruebas de comparaciones múltiples sobre la posición de la caldera. Explique qué es lo que muestran los resultados.

14.11 Se realizó un estudio para determinar la influencia de 2 factores, el método de análisis y el laboratorio que hace el análisis, sobre el nivel de contenido de azufre del carbón. Se asignaron al azar 28 especímenes de carbón a 14 combinaciones de factores, la estructura de las unidades experimentales representada por las combinaciones de 7 laboratorios y 2 métodos de análisis con 2 especímenes por combinación de factores. Los datos, expresados en porcentaje de azufre, son los siguientes.

Laboratorio	**Método**			
	1		**2**	
1	0.109	0.105	0.105	0.108
2	0.129	0.122	0.127	0.124
3	0.115	0.112	0.109	0.111
4	0.108	0.108	0.117	0.118
5	0.097	0.096	0.110	0.097
6	0.114	0.119	0.116	0.122
7	0.155	0.145	0.164	0.160

(Los datos se tomaron de G. Taguchi, "Signal to Noise Ratio and Its Applications to Testing Material", *Reports of Statistical Application Research*, Union of Japanese Scientists and Engineers, Vol. 18, Núm. 4, 1971).

a) Haga un análisis de varianza y exprese los resultados en la tabla correspondiente.
b) ¿Es significativa la interacción? Si lo es, analice lo que significa para el científico. Utilice un valor *P* en sus conclusiones.

c) ¿Son estadísticamente significativos los efectos principales individuales, el laboratorio y el método de análisis? Analice la información y lo que aprendió y base su respuesta en el contexto de cualquier interacción significativa.
d) Dibuje una gráfica de interacción que ilustre el efecto de la interacción.
e) Efectúe una prueba para comparar los métodos 1 y 2 en el laboratorio 1, y haga lo mismo para el laboratorio 7. Comente lo que revelan esos resultados.

14.12 En un experimento efectuado en el departamento de Ingeniería Civil de Virginia Tech se observó el crecimiento que cierto tipo de alga tenía en el agua, en función del tiempo y la dosis de cobre que se agregaba al líquido. Los datos se presentan a continuación. La respuesta se expresa en unidades de algas.

	Tiempo en días		
Cobre	**5**	**12**	**18**
1	0.30	0.37	0.25
	0.34	0.36	0.23
	0.32	0.35	0.24
2	0.24	0.30	0.27
	0.23	0.32	0.25
	0.22	0.31	0.25
3	0.20	0.30	0.27
	0.28	0.31	0.29
	0.24	0.30	0.25

a) Haga un análisis de varianza y elabore la tabla correspondiente.
b) Comente acerca de si los datos son suficientes para mostrar un efecto del tiempo sobre la concentración de algas.
c) Haga lo mismo para el contenido de cobre. ¿El nivel de contenido de cobre tiene algún efecto sobre la concentración de algas?
d) Comente los resultados de la prueba de interacción. ¿Cómo influye el tiempo sobre el efecto del contenido de cobre?

14.13 En *Classical and Modern Regression with Applications* (Duxbury Classic Series, 2a. ed., 1990), de Myers, se describe un experimento en el que la Agencia de Protección Ambiental busca determinar el efecto de 2 métodos de tratamiento de aguas sobre la absorción del magnesio. Se miden los niveles de magnesio, en gramos por centímetro cúbico (cc) y se incorporan 2 niveles diferentes de tiempo al experimento. Los datos son los siguientes:

	Tratamiento					
Tiempo (horas)	**1**			**2**		
1	2.19	2.15	2.16	2.03	2.01	2.04
2	2.01	2.03	2.04	1.88	1.86	1.91

a) Dibuje una gráfica de la interacción. ¿Cuál es su impresión?
b) Efectúe un análisis de varianza y presente pruebas para los efectos principales y de interacción.
c) Mencione los hallazgos científicos acerca de cómo influyen el tiempo y el tratamiento en la absorción del magnesio.
d) Ajuste el modelo de regresión adecuado usando el tratamiento como variable categórica. Incluya la interacción en el modelo.
e) ¿La interacción es significativa en el modelo de regresión?

14.14 Considere los datos del ejercicio 14.12 y responda las siguientes preguntas.
a) Ambos factores, el cobre y el tiempo, son cuantitativos. Como resultado, podría ser de interés un modelo de regresión. Describa cuál sería un modelo adecuado si se usa x_1 = contenido de cobre y x_2 = tiempo. Ajuste el modelo a los datos mostrando los coeficientes de regresión y haga una prueba *t* para cada uno.
b) Ajuste el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon,$$

y compárelo con el que eligió en el inciso *a*. ¿Cuál es más apropiado? Como criterio utilice R^2_{ajus}.

14.15 El propósito del estudio *The Incorporation of a Chelating Agent into a Flame Retardant Finish of a Cotton Flannelette and the Evaluation of Selected Fabric Properties*, llevado a cabo en Virginia Tech, fue evaluar el uso de un agente quelante como parte del acabado retardante del fuego de la franela de algodón, determinando sus efectos en la inflamabilidad después de lavar la tela en condiciones específicas. Se utilizaron 2 tratamientos con 2 niveles; se prepararon 2 baños, uno con celulosa de carboximetilo (baño I) y otro sin ella (baño II). La mitad de la tela se lavó 5 veces y la otra mitad se lavó 10 veces. Hubo 12 pedazos de tela en cada combinación de baño/número de lavados. Después de los lavados se midieron las longitudes quemadas de la tela, así como los tiempos de combustión. Se registraron los siguientes tiempos de combustión (en segundos):

Lavados	**Baño I**			**Baño II**		
5	13.7	23.0	15.7	6.2	5.4	5.0
	25.5	15.8	14.8	4.4	5.0	3.3
	14.0	29.4	9.7	16.0	2.5	1.6
	14.0	12.3	12.3	3.9	2.5	7.1
10	27.2	16.8	12.9	18.2	8.8	14.5
	14.9	17.1	13.0	14.7	17.1	13.9
	10.8	13.5	25.5	10.6	5.8	7.3
	14.2	27.4	11.5	17.7	18.3	9.9

a) Realice un análisis de varianza. ¿Existe un término de interacción significativo?
b) ¿Se encontraron diferencias en los efectos principales? Analice la información.

14.4 Experimentos de tres factores

En esta sección consideramos un experimento con 3 factores, A, B y C, en los niveles a, b y c, respectivamente, en un diseño experimental completamente aleatorizado. Suponga de nuevo que se tienen n observaciones para cada una de las abc combinaciones de tratamientos. Debemos proceder a realizar las pruebas de significancia para los 3 efectos principales y las interacciones implicadas. Se espera que el lector podrá utilizar después esta descripción para generalizar el análisis a $k > 3$ factores.

Modelo para el experimento de tres factores

El modelo para el experimento de 3 factores es

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

$i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, c$; y $l = 1, 2, \ldots, n$, donde α_i, β_j y γ_k son los efectos principales y $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ y $(\beta\gamma)_{jk}$ son los efectos de la interacción de 2 factores que tienen la misma interpretación que en el experimento con 2 factores.

El término $(\alpha\beta\gamma)_{ijk}$ se denomina **efecto de interacción de 3 factores**, y representa la no aditividad de las $(\alpha\beta)_{ij}$ sobre los diferentes niveles del factor C. Igual que antes, la suma de todos los efectos principales es igual a 0, y la suma sobre cualesquiera de los subíndices de los efectos de la interacción entre 2 y 3 factores es igual a 0. En muchas situaciones experimentales estas interacciones de orden superior son insignificantes y sus cuadrados medios sólo reflejan variación aleatoria; pero se debe describir el análisis en su forma más general.

Nuevamente, para realizar pruebas válidas de significancia debe suponerse que los errores son valores de variables aleatorias independientes y con distribución normal, cada una con media igual a 0 y varianza común σ^2.

La filosofía general respecto al análisis es la misma que la que se estudió para los experimentos de 1 y 2 factores. La suma de cuadrados se divide en 8 términos, donde cada uno representa una fuente de variación de los que se obtienen estimados independientes de σ^2 cuando todos los efectos principales y de la interacción son iguales a 0. Si los efectos de cualquier factor dado o interacción no son iguales a 0, entonces el cuadrado medio estimará la varianza del error más un componente debido al efecto sistemático en cuestión.

Suma de cuadrados para un experimento de tres factores

$$SCA = bcn \sum_{i=1}^{a} (\bar{y}_{i...} - \bar{y}_{....})^2 \qquad SC(AB) = cn \sum_i \sum_j (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2$$

$$SCB = acn \sum_{j=1}^{b} (\bar{y}_{.j..} - \bar{y}_{....})^2 \qquad SC(AC) = bn \sum_i \sum_k (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$SCC = abn \sum_{k=1}^{c} (\bar{y}_{..k.} - \bar{y}_{....})^2 \qquad SC(BC) = an \sum_j \sum_k (\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$SC(ABC) = n \sum_i \sum_j \sum_k (\bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} - \bar{y}_{....})^2$$

$$STC = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{....})^2 \qquad SCE = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{ijk.})^2$$

Aunque en esta sección hacemos énfasis en la interpretación de una salida de resultados por computadora con comentarios, en vez de preocuparnos por cálculos laboriosos de sumas de cuadrados, ofrecemos lo siguiente como la suma de cuadrados para los 3 efectos principales y las interacciones. Observe la evidente extensión del problema de 2 factores a uno de 3.

Los promedios en las fórmulas se definen como sigue:

$\bar{y}_{...}$ = promedio de todas las *abcn* observaciones,
$\bar{y}_{i...}$ = promedio de las observaciones para el *i*-ésimo nivel del factor *A*,
$\bar{y}_{.j.}$ = promedio de las observaciones para el *j*-ésimo nivel del factor *B*,
$\bar{y}_{..k.}$ = promedio de las observaciones para el *k*-ésimo nivel del factor *C*,
$\bar{y}_{ij.}$ = promedio de las observaciones para el *i*-ésimo nivel de *A* y el *j*-ésimo nivel de *B*,
$\bar{y}_{i.k.}$ = promedio de las observaciones para el *i*-ésimo nivel de *A* y el *k*-ésimo nivel de *C*,
$\bar{y}_{.jk.}$ = promedio de las observaciones para el *j*-ésimo nivel de *B* y el *k*-ésimo nivel de *C*,
$\bar{y}_{ijk.}$ = promedio de las observaciones para la (*ijk*)-ésima combinación de tratamientos.

Los cálculos en una tabla de análisis de varianza para un problema de 3 factores con *n* réplicas de corridas para cada combinación de factores se resumen en la tabla 14.7.

Tabla 14.7: ANOVA para el experimento de 3 factores con *n* réplicas

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**
Efecto principal:				
A	*SCA*	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	*SCB*	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	*SCC*	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de 2 factores:				
AB	*SC*(*AB*)	$(a-1)(b-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
AC	*SC*(*AC*)	$(a-1)(c-1)$	s_5^2	$f_5 = \frac{s_5^2}{s^2}$
BC	*SC*(*BC*)	$(b-1)(c-1)$	s_6^2	$f_6 = \frac{s_6^2}{s^2}$
Interacción de 3 factores:				
ABC	*SC*(*ABC*)	$(a-1)(b-1)(c-1)$	s_7^2	$f_7 = \frac{s_7^2}{s^2}$
Error	*SCE*	$abc(n-1)$	s^2	
Total	*STC*	$abcn-1$		

Para el experimento de 3 factores con una sola corrida experimental por combinación se podría utilizar el análisis de la tabla 14.7 con $n = 1$ y usando la suma de cuadrados de la interacción *ABC* para *SCE*. En este caso suponemos que los efectos de la interacción $(\alpha\beta\gamma)_{ijk}$ son todos iguales a cero, de modo que

$$E\left[\frac{SC(ABC)}{(a-1)(b-1)(c-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)(c-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\alpha\beta\gamma)_{ijk}^2 = \sigma^2.$$

Es decir, $SC(ABC)$ representa la variación que sólo se debe al error experimental. Por lo tanto, su cuadrado medio proporciona un estimado no sesgado de la varianza del error. Con $n = 1$ y $SCE = SC(ABC)$, la suma de cuadrados del error se obtiene restando la suma de cuadrados de los efectos principales y las interacciones de 2 factores a la suma de cuadrados total.

Ejemplo 14.4: En la producción de un material en particular hay 3 variables de interés: A, el efecto del operador (3 operadores): B, el catalizador utilizado en el experimento (3 catalizadores); y C, el tiempo de lavado del producto después del proceso de enfriamiento (15 minutos y 20 minutos). Se realizaron 3 corridas con cada combinación de factores. Se consideró que debían estudiarse todas las interacciones entre los factores. En la tabla 14.8 se presentan los productos codificados. Realice un análisis de varianza para probar si existen efectos significativos.

Tabla 14.8: Datos para el ejemplo 14.4

	Tiempo de lavado, *C*					
	15 minutos			**20 minutos**		
	Catalizador, *B*			**Catalizador, *B***		
Operador, *A*	**1**	**2**	**3**	**1**	**2**	**3**
1	10.7	10.3	11.2	10.9	10.5	12.2
	10.8	10.2	11.6	12.1	11.1	11.7
	11.3	10.5	12.0	11.5	10.3	11.0
2	11.4	10.2	10.7	9.8	12.6	10.8
	11.8	10.9	10.5	11.3	7.5	10.2
	11.5	10.5	10.2	10.9	9.9	11.5
3	13.6	12.0	11.1	10.7	10.2	11.9
	14.1	11.6	11.0	11.7	11.5	11.6
	14.5	11.5	11.5	12.7	10.9	12.2

***Solución*:** La tabla 14.9 muestra el análisis de varianza de los datos. Ninguna de las interacciones muestra un efecto significativo a un nivel $\alpha = 0.05$. Sin embargo, el valor P para BC es 0.0610, de modo que no debe ignorarse. Los efectos del operador y el catalizador son significativos, en tanto que el del tiempo de lavado no lo es. ∎

Impacto de la interacción *BC*

Se deben analizar otros aspectos del ejemplo 14.4, en particular acerca del manejo del efecto que la interacción entre el catalizador y el tiempo de lavado tienen sobre la prueba del efecto principal del tiempo de lavado (factor C). Recuerde el análisis de la sección 14.2. Se proporcionaron ejemplos de la manera en que la presencia de la interacción podría cambiar la interpretación que se da a los efectos principales. En el ejemplo 14.4 la interacción BC es significativa aproximadamente al nivel 0.06. No obstante, suponga que se obtiene una tabla de medias de 2 factores como la 14.10.

Queda claro por qué el tiempo de lavado no fue significativo. Un analista poco cuidadoso se quedaría con la impresión de que el tiempo de lavado podría eliminarse de cualquier estudio futuro en el que se mida el producto. Sin embargo, es notorio cómo cambia el efecto del tiempo de lavado de uno negativo para el primer catalizador, a lo

Tabla 14.9: ANOVA para un experimento de 3 factores en un diseño completamente aleatorizado

Fuente	gl	Suma de cuadrados	Cuadrado medio	Valor *F*	Valor *P*
A	2	13.98	6.99	11.64	0.0001
B	2	10.18	5.09	8.48	0.0010
AB	4	4.77	1.19	1.99	0.1172
C	1	1.19	1.19	1.97	0.1686
AC	2	2.91	1.46	2.43	0.1027
BC	2	3.63	1.82	3.03	0.0610
ABC	4	4.91	1.23	2.04	0.1089
Error	36	21.61	0.60		
Total	53	63.19			

Tabla 14.10: Tabla de medias de 2 factores para el ejemplo 14.4

	Tiempo de lavado, *C*	
Catalizador, *B*	**15 min**	**20 min**
1	12.19	11.29
2	10.86	10.50
3	11.09	11.46
Medias	11.38	11.08

que parece ser un efecto positivo para el tercer catalizador. Si sólo nos concentramos en los datos para el catalizador 1, una comparación simple entre las medias de los 2 tiempos de lavado produciría un estadístico *t* sencillo:

$$t = \frac{12.19 - 11.29}{\sqrt{0.6(2/9)}} = 2.5,$$

que es significativo a un nivel menor que 0.02. Así, bien podría ignorarse un importante efecto negativo del tiempo de lavado para el catalizador 1 si el analista hace la interpretación general incorrecta de la razón *F* insignificante del tiempo de lavado.

Agrupamiento en modelos multifactoriales

El modelo de 3 factores y su análisis se describió de la manera más general mediante la inclusión en el modelo de todas las interacciones posibles. Por supuesto, hay muchas situaciones en las que *a priori* se sabe que el modelo no debería contener ciertas interacciones. Así, es posible aprovechar este conocimiento al combinar o agrupar las sumas de cuadrados correspondientes a interacciones despreciables con la suma de cuadrados del error para formar un nuevo estimador de σ^2 con un número más grande de grados de libertad. Por ejemplo, en un experimento de metalurgia diseñado para estudiar el efecto de 3 variables importantes del proceso sobre el espesor de película, suponga que se sabe que el factor *A*, la concentración de ácido, no interactúa con los factores *B* y *C*. Las

Tabla 14.11: ANOVA sin interacción del factor A

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	**f calculada**
Efecto principal:				
A	SCA	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	SCB	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	SCC	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de 2 factores:				
BC	$SC(BC)$	$(b-1)(c-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
Error	SCE	Resta	s^2	
Total	STC	$abcn-1$		

sumas de cuadrados SCA, SCB, SCC y $SC(BC)$ se calculan usando los métodos descritos en un apartado anterior de esta sección. Todos los cuadrados medios de los efectos restantes ahora estimarán de manera independiente la varianza del error σ^2. Por lo tanto, formamos el nuevo **cuadrado medio del error agrupando** $SC(AB)$, $SC(AC)$, $SC(ABC)$ y SCE junto con los grados de libertad correspondientes. El denominador resultante de las pruebas de significancia es, entonces, el cuadrado medio del error dado por

$$s^2 = \frac{SC(AB) + SC(AC) + SC(ABC) + SCE}{(a-1)(b-1) + (a-1)(c-1) + (a-1)(b-1)(c-1) + abc(n-1)}.$$

Por supuesto, con una resta se obtienen la suma de cuadrados agrupada y los grados de libertad agrupados, una vez que se calcula la STC y las sumas de cuadrados para los efectos existentes. La tabla del análisis de varianza adoptaría así la forma de la tabla 14.11.

Experimentos factoriales en bloques

En este capítulo se ha supuesto que el diseño experimental utilizado es un diseño completamente aleatorizado. Al interpretar los niveles del factor A en la tabla 14.11 **como bloques diferentes** se tiene el procedimiento del análisis de varianza para un experimento de 2 factores en un diseño de bloques aleatorizados. Por ejemplo, si se interpretan los operadores del ejemplo 14.4 como bloques, y se supone que no hay interacción entre los bloques y los otros 2 factores, el análisis de varianza adopta la forma de la tabla 14.12, en vez de la de la tabla 14.9. El lector puede verificar que el cuadrado medio del error también es

$$s^2 = \frac{4.77 + 2.91 + 4.91 + 21.61}{4 + 2 + 4 + 36} = 0.74,$$

lo que demuestra el agrupamiento de las sumas de cuadrados para los efectos de la interacción inexistente. Observe que el factor B, el catalizador, tiene un efecto significativo sobre el producto.

Tabla 14.12: ANOVA para un experimento de 2 factores en un diseño de bloques aleatorizados

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*f* calculada	Valor *P*
Bloques	13.98	2	6.99		
Efecto principal:					
B	10.18	2	5.09	6.88	0.0024
C	1.18	1	1.18	1.59	0.2130
Interacción de 2 factores					
BC	3.64	2	1.82	2.46	0.0966
Error	34.21	46	0.74		
Total	63.19	53			

Ejemplo 14.5: Se realizó un experimento para determinar los efectos de la temperatura, la presión y la intensidad de agitación sobre la tasa de filtración del producto. Esto se hizo en una planta piloto. El experimento se corrió en 2 niveles de cada factor. Además, se decidió que debían utilizarse 2 lotes de materia prima, los cuales fueron tratados como bloques. Se hicieron 8 corridas experimentales en orden aleatorio para cada lote de materia prima. Se piensa que todas las interacciones de los 2 factores podrían ser de interés. No se supone que haya interacciones con los lotes. Los datos aparecen en la tabla 14.13. Las letras "B" y "A" implican niveles bajo y alto, respectivamente. La tasa de filtración se expresa en galones por hora.

a) Elabore la tabla ANOVA completa. Agrupe todas las "interacciones" con los bloques dentro del error.

b) ¿Cuáles interacciones parecen ser significativas?

c) Construya gráficas que revelen las interacciones significativas e interprételas. Explique el significado de la gráfica para el ingeniero.

Tabla 14.13: Datos para el ejemplo 14.5

Lote 1					
	Tasa de agitación baja			**Tasa de agitación alta**	
Temp.	**Presión B**	**Presión A**	**Temp.**	**Presión B**	**Presión A**
B	43	49	**B**	44	47
A	64	68	**A**	97	102
Lote 2					
	Tasa de agitación baja			**Tasa de agitación alta**	
Temp.	**Presión B**	**Presión A**	**Temp.**	**Presión B**	**Presión A**
B	49	57	**B**	51	55
A	70	76	**A**	103	106

***Solución*:** *a*) En la figura 14.7 se presenta una salida de resultados impresos por computadora del *SAS*.

b) Como se aprecia en la figura 14.7, la interacción de la temperatura con la tasa de agitación (*strate*) parece ser muy significativa. Asimismo, la interacción de la presión con la tasa de agitación también parece ser significativa. A propósito, si se hicieran más agrupamientos al combinar las interacciones insignificantes con el error, las conclusiones serían las mismas y el valor *P* para la interacción de la presión con la tasa de agitación se volvería más fuerte, a saber, 0.0517.

c) Como se aprecia en la figura 14.7, los efectos principales tanto de la tasa de agitación como de la temperatura son muy significativos. Un vistazo a la gráfica de interacción de la figura 14.8*a* revela que el efecto de la tasa de agitación depende del nivel de la temperatura. Con la temperatura baja el efecto de la tasa de agitación es despreciable, mientras que con la temperatura alta la tasa de agitación tiene un efecto positivo fuerte sobre la tasa media de filtración. En la figura 14.8*b* la interacción entre la presión y la tasa de agitación, aunque no de manera tan pronunciada como la de la figura 14.8*a*, todavía muestra una ligera inconsistencia del efecto de la tasa de agitación a través de la presión.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
batch	1	175.562500	175.562500	177.14	<.0001
pressure	1	95.062500	95.062500	95.92	<.0001
temp	1	5292.562500	5292.562500	5340.24	<.0001
pressure*temp	1	0.562500	0.562500	0.57	0.4758
strate	1	1040.062500	1040.062500	1049.43	<.0001
pressure*strate	1	5.062500	5.062500	5.11	0.0583
temp*strate	1	1072.562500	1072.562500	1082.23	<.0001
pressure*temp*strate	1	1.562500	1.562500	1.58	0.2495
Error	7	6.937500	0.991071		
Corrected Total	15	7689.937500			

Figura 14.7: ANOVA para el ejemplo 14.5, interacción del lote agrupado con el error.

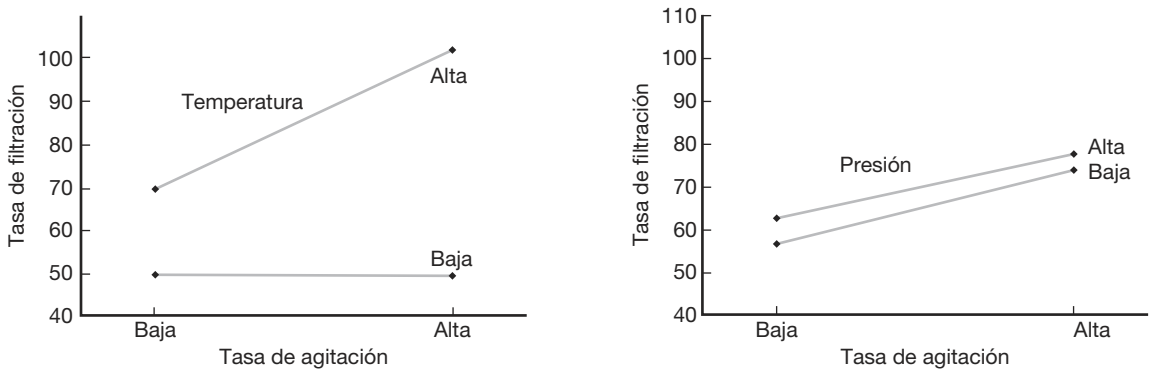


Figura 14.8: Gráficas de interacción para el ejemplo 14.5.

Ejercicios

14.16 Considere una situación experimental que implique los factores A, B y C, en la que se supone un modelo de efectos fijos de 3 factores de la forma $y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \epsilon_{ijkl}$. Se considera que todas las demás interacciones no existen o son despreciables. Los datos se presentan en seguida.

	B_1			B_2		
	C_1	C_2	C_3	C_1	C_2	C_3
A_1	4.0	3.4	3.9	4.4	3.1	3.1
	4.9	4.1	4.3	3.4	3.5	3.7
A_2	3.6	2.8	3.1	2.7	2.9	3.7
	3.9	3.2	3.5	3.0	3.2	4.2
A_3	4.8	3.3	3.6	3.6	2.9	2.9
	3.7	3.8	4.2	3.8	3.3	3.5
A_4	3.6	3.2	3.2	2.2	2.9	3.6
	3.9	2.8	3.4	3.5	3.2	4.3

a) Haga una prueba de significancia sobre la interacción BC al nivel $\alpha = 0.05$.

b) Desarrolle pruebas de significancia sobre los efectos principales A, B y C usando un cuadrado medio del error agrupado, con un nivel $\alpha = 0.05$.

14.17 Los siguientes datos son medidas de un experimento donde se usaron 3 factores, A, B y C, todos de efectos fijos.

	C_1			C_2			C_3		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
A_1	15.0	14.8	15.9	16.8	14.2	13.2	15.8	15.5	19.2
	18.5	13.6	14.8	15.4	12.9	11.6	14.3	13.7	13.5
	22.1	12.2	13.6	14.3	13.0	10.1	13.0	12.6	11.1
A_2	11.3	17.2	16.1	18.9	15.4	12.4	12.7	17.3	7.8
	14.6	15.5	14.7	17.3	17.0	13.6	14.2	15.8	11.5
	18.2	14.2	13.4	16.1	18.6	15.2	15.9	14.6	12.2

a) Haga pruebas de significancia sobre todas las interacciones a un nivel $\alpha = 0.05$.

b) Realice pruebas de significancia sobre los efectos principales a un nivel $\alpha = 0.05$.

c) Dé una explicación de la forma en que una interacción significativa enmascara el efecto del factor C.

14.18 El método de fluorescencia por rayos X es una herramienta analítica importante para determinar la concentración de material en los propulsores sólidos para misiles. En el artículo *An X-ray Fluorescence Method for Analyzing Polybutadiene Acrylic Acid (PBAA) Propellants* (Quarterly Report, RK-TR-62-1, Army Ordinance Missile Command, 1962), se afirma que el proceso de mezcla del propulsor y el tiempo de análisis influyen en la homogeneidad del material y, por lo tanto, en la precisión de las mediciones de la intensidad de los rayos X. Se hizo un experimento utilizando 3 factores: A, las condiciones de mezcla (4 niveles); B, el tiempo de análisis (2 niveles); y C, el método de carga del propulsor en los recipientes para muestras (temperatura elevada y de la habitación). Se obtuvieron los datos siguientes, que representan el porcentaje de peso del perclorato de amoniaco en un propulsor dado.

	Método de carga, C			
	Caliente		**Temp. de la hab.**	
A	B_1	B_2	B_1	B_2
1	38.62	38.45	39.82	39.82
	37.20	38.64	39.15	40.26
	38.02	38.75	39.78	39.72
2	37.67	37.81	39.53	39.56
	37.57	37.75	39.76	39.25
	37.85	37.91	39.90	39.04
3	37.51	37.21	39.34	39.74
	37.74	37.42	39.60	39.49
	37.58	37.79	39.62	39.45
4	37.52	37.60	40.09	39.36
	37.15	37.55	39.63	39.38
	37.51	37.91	39.67	39.00

a) Realice un análisis de varianza con $\alpha = 0.01$ para probar la existencia de efectos principales y de interacción significativos.

b) Analice la influencia de los 3 factores sobre el porcentaje de peso del perclorato de amoniaco. Incluya en su análisis el papel que desempeña cualquier interacción significativa.

14.19 La fatiga por corrosión de los metales se ha definido como la acción simultánea de tensión cíclica y ataque químico sobre una estructura metálica. En el estudio *Effect of Humidity and Several Surface Coatings on the Fatigue Life of 2024-T351 Aluminum Alloy*, realizado por el Departamento de Ingeniería Mecánica de Virginia Tech, se utilizó una técnica que requería la aplicación de un recubrimiento protector de cromato para minimizar el daño de la fatiga por corrosión en el aluminio. En la investigación se emplearon 3 factores con 5 réplicas para cada combinación de tratamientos: recubrimiento, en 2 niveles; humedad y esfuerzo cortante, ambos en 3 niveles. A continuación se presentan los datos de fatiga expresados en miles de ciclos antes de la falla.

a) Realice un análisis de varianza con $\alpha = 0.01$ para probar la existencia de efectos principales y de interacción significativos.

b) Haga una recomendación para las combinaciones de los 3 factores que producirían poco daño por fatiga.

Recubrimiento	Humedad	Esfuerzo cortante (psi) 13,000	17,000	20,000
Sin recubrimiento	**Bajo (20–25% RH)**	4580	5252	361
		10,126	897	466
		1341	1465	1069
		6414	2694	469
		3549	1017	937
	Medio (50–60% RH)	2858	799	314
		8829	3471	244
		10,914	685	261
		4067	810	522
		2595	3409	739
	Alto (86–91% RH)	6489	1862	1344
		5248	2710	1027
		6816	2632	663
		5860	2131	1216
		5901	2470	1097
Cromado	**Bajo (20–25% RH)**	5395	4035	130
		2768	2022	841
		1821	914	1595
		3604	2036	1482
		4106	3524	529
	Medio (50–60% RH)	4833	1847	252
		7414	1684	105
		10,022	3042	847
		7463	4482	874
		21,906	996	755
	Alto (86–91% RH)	3287	1319	586
		5200	929	402
		5493	1263	846
		4145	2236	524
		3336	1392	751

14.20 Para un estudio de la dureza de los empastes dentales de oro se eligieron 5 dentistas al azar y se asignaron a combinaciones de 3 métodos de condensación y 2 tipos de oro. Se midió la dureza. (Véase Hoaglin, Mosteller y Tukey, 1991). Permita que los dentistas desempeñen el papel de bloques. Los datos se presentan a continuación.

a) Proponga el modelo adecuado con las suposiciones.

b) ¿Hay una interacción significativa entre el método de condensación y el tipo de material de empaste de oro?

c) ¿Hay un método de condensación que parezca mejor? Explique su respuesta.

Dentista	Método	Tipo Lámina dorada	Goldent
1	**1**	792	824
	2	772	772
	3	782	803
2	**1**	803	803
	2	752	772
	3	715	707
3	**1**	715	724
	2	792	715
	3	762	606
4	**1**	673	946
	2	657	743
	3	690	245
5	**1**	634	715
	2	649	724
	3	724	627

(*cont.*)

14.21 Las copiadoras electrónicas funcionan adhiriendo tinta negra al papel mediante electricidad estática. La etapa final del proceso de copiado comprende el calentamiento y adhesión de la tinta sobre el papel. La potencia de la adhesión durante este proceso final determina la calidad de la copia. Se plantea que la temperatura, el estado superficial de la adhesión en el rodillo y la dureza del rodillo de la prensa influyen en la potencia de adhesión de la copiadora. Se hizo un experimento con tratamientos, que consistían en una combinación de estos 3 factores en cada uno de 3 niveles. Los datos siguientes muestran la potencia de la adhesión para cada combinación de tratamientos. Lleve a cabo un análisis de varianza con $\alpha = 0.05$ para probar si hay efectos principales y de interacción significativos.

	Estado superficial de la adhesión en el rodillo	Dureza del rodillo de la prensa 20		40		60	
Temp. baja	**Suave**	0.52	0.44	0.54	0.52	0.60	0.55
		0.57	0.53	0.65	0.56	0.78	0.68
	Medio	0.64	0.59	0.79	0.73	0.49	0.48
		0.58	0.64	0.79	0.78	0.74	0.50
	Duro	0.67	0.77	0.58	0.68	0.55	0.65
		0.74	0.65	0.57	0.59	0.57	0.58
Temp. media	**Suave**	0.46	0.40	0.31	0.49	0.56	0.42
		0.58	0.37	0.48	0.66	0.49	0.49
	Medio	0.60	0.43	0.66	0.57	0.64	0.54
		0.62	0.61	0.72	0.56	0.74	0.56
	Duro	0.53	0.65	0.53	0.45	0.56	0.66
		0.66	0.56	0.59	0.47	0.71	0.67
Temp. alta	**Suave**	0.52	0.44	0.54	0.52	0.65	0.49
		0.57	0.53	0.65	0.56	0.65	0.52
	Medio	0.53	0.65	0.53	0.45	0.49	0.48
		0.66	0.56	0.59	0.47	0.74	0.50
	Duro	0.43	0.43	0.48	0.31	0.55	0.65
		0.47	0.44	0.43	0.27	0.57	0.58

14.22 Considere el conjunto de datos del ejercicio 14.21.

a) Construya una gráfica de la interacción para cualquier interacción de 2 factores que sea significativa.

b) Dibuje una gráfica de probabilidad normal de residuales y coméntela.

14.23 Considere combinaciones de 3 factores en el retiro de la suciedad de cargas estándar de lavandería. El primer factor es la marca del detergente: *X*, *Y* o *Z*. El segundo factor es el tipo de detergente: líquido o en polvo. El tercer factor es la temperatura del agua, caliente o tibia. El experimento se replicó 3 veces. La respuesta es el porcentaje de suciedad eliminada. Los datos son los siguientes:

Marca	Tipo	Temperatura			
X	**En polvo**	**Caliente**	85	88	80
		Tibia	82	83	85
	Líquido	**Caliente**	78	75	72
		Tibia	75	75	73
Y	**En polvo**	**Caliente**	90	92	92
		Tibia	88	86	88
	Líquido	**Caliente**	78	76	70
		Tibia	76	77	76
Z	**En polvo**	**Caliente**	85	87	88
		Tibia	76	74	78
	Líquido	**Caliente**	60	70	68
		Tibia	55	57	54

a) ¿Existen efectos de la interacción significativos a un nivel $\alpha = 0.05$?
b) ¿Hay diferencias significativas entre las tres marcas de detergente?
c) ¿Cuál combinación de factores preferiría utilizar?

14.24 Un científico recaba datos experimentales sobre el radio de un grano de combustible propulsor, *y*, en función de la temperatura del polvo, la tasa de extrusión y la temperatura del molde. Los resultados de los 3 factores del experimento son los siguientes:

	Temp. del polvo			
	150		190	
	Temp. del molde		Temp. del molde	
Tasa	220	250	220	250
12	82	124	88	129
24	114	157	121	164

No se dispone de recursos para hacer experimentos repetidos con las 8 combinaciones de factores. Se cree que la tasa de extrusión no interactúa con la temperatura del molde, y que la interacción entre los 3 factores es despreciable. Así, esas 2 interacciones pueden agruparse para producir un término de "error" con 2 grados de libertad.

a) Haga un análisis de varianza que incluya los 3 efectos principales e interacciones de 2 factores. Determine cuáles efectos influyen en el radio del grano de combustible.
b) Construya gráficas de interacción para la temperatura del polvo usando la temperatura del molde y la del polvo mediante las interacciones de la tasa de extrusión.
c) Comente acerca de la consistencia de la apariencia de las gráficas de interacción y las pruebas sobre las 2 interacciones en el ANOVA.

14.25 En el libro *Design of Experiments for Quality Improvement*, publicado por la Japanese Standards Association (1989), se reporta un estudio sobre la extracción de polietileno por medio de un solvente, y la manera en que la cantidad de gel (proporción) se ve influida por 3 factores: el tipo de solvente, la temperatura de extracción y el tiempo de extracción. Se diseñó un experimento factorial y se obtuvieron los datos siguientes, expresados en proporción de gel.

		Tiempo					
Temp. del solvente		4		8		16	
Etanol	**120**	94.0	94.0	93.8	94.2	91.1	90.5
	80	95.3	95.1	94.9	95.3	92.5	92.4
Tolueno	**120**	94.6	94.5	93.6	94.1	91.1	91.0
	80	95.4	95.4	95.6	96.0	92.1	92.1

a) Haga un análisis de varianza y determine cuáles factores e interacciones influyen en la proporción de gel.
b) Construya una gráfica de la interacción entre cualesquiera 2 factores que sea significativa. Además, explique qué conclusión se podría extraer de la presencia de la interacción.
c) Haga una gráfica de probabilidad normal de los residuales y comente.

14.5 Experimentos factoriales para efectos aleatorios y modelos mixtos

En un experimento de 2 factores con efectos aleatorios se tiene el modelo

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk},$$

para $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; y $k = 1, 2, \ldots, n$, donde A_i, B_j $(AB)_{ij}$ y ϵ_{ijk} son variables aleatorias independientes con medias igual a 0 y varianzas σ_α^2, σ_β^2, $\sigma_{\alpha\beta}^2$ y σ^2, respectivamente. Las sumas de cuadrados para experimentos de efectos aleatorios se calculan

exactamente de la misma forma que en los experimentos de efectos fijos. Ahora se tiene interés en probar hipótesis con la forma

$$H_0': \sigma_\alpha^2 = 0, \qquad H_0'': \sigma_\beta^2 = 0, \qquad H_0''': \sigma_{\alpha\beta}^2 = 0,$$
$$H_1': \sigma_\alpha^2 \neq 0, \qquad H_1'': \sigma_\beta^2 \neq 0, \qquad H_1''': \sigma_{\alpha\beta}^2 \neq 0,$$

donde el denominador en la razón f no es necesariamente el cuadrado medio del error. El denominador apropiado se determina examinando los valores esperados de los distintos cuadrados medios, los cuales se presentan en la tabla 14.14.

Tabla 14.14: Cuadrados medios esperados para un experimento de efectos aleatorios de 2 factores

Fuente de variación	Grados de libertad	Cuadrado medio	Cuadrado medio esperado
A	$a-1$	s_1^2	$\sigma^2 + n\sigma_{\alpha\beta}^2 + bn\sigma_\alpha^2$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma_{\alpha\beta}^2 + an\sigma_\beta^2$
AB	$(a-1)(b-1)$	s_3^2	$\sigma^2 + n\sigma_{\alpha\beta}^2$
Error	$ab(n-1)$	s^2	σ^2
Total	$abn-1$		

En la tabla 14.14 se observa que H_0' y H_0'' se prueban usando s_3^2 en el denominador de la razón f; mientras que H_0''' se prueba con s^2 en el denominador. Los estimados no sesgados de los componentes de la varianza son

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_{\alpha\beta}^2 = \frac{s_3^2 - s^2}{n}, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s_3^2}{bn}, \qquad \hat{\sigma}_\beta^2 = \frac{s_2^2 - s_3^2}{an}.$$

Tabla 14.15: Cuadrados medios esperados para un experimento de efectos aleatorios de 3 factores

Fuente de variación	Grados de libertad	Cuadrado medio	Cuadrado medio esperado
A	$a-1$	s_1^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + cn\sigma_{\alpha\beta}^2 + bn\sigma_{\alpha\gamma}^2 + bcn\sigma_\alpha^2$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + cn\sigma_{\alpha\beta}^2 + an\sigma_{\beta\gamma}^2 + acn\sigma_\beta^2$
C	$c-1$	s_3^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + bn\sigma_{\alpha\gamma}^2 + an\sigma_{\beta\gamma}^2 + abn\sigma_\gamma^2$
AB	$(a-1)(b-1)$	s_4^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + cn\sigma_{\alpha\beta}^2$
AC	$(a-1)(c-1)$	s_5^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + bn\sigma_{\alpha\gamma}^2$
BC	$(b-1)(c-1)$	s_6^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + an\sigma_{\beta\gamma}^2$
ABC	$(a-1)(b-1)(c-1)$	s_7^2	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2$
Error	$abc(n-1)$	s^2	σ^2
Total	$abcn-1$		

En la tabla 14.15 se presentan los cuadrados medios esperados para el experimento de 3 factores con efectos aleatorios en un diseño completamente aleatorizado. A partir de los cuadrados medios esperados de la tabla 14.15 es evidente que se pueden formar

razones f adecuadas para probar todos los componentes de la varianza de la interacción de 2 y 3 factores. Sin embargo, para probar una hipótesis de la forma

$$H_0: \ \sigma_\alpha^2 = 0,$$
$$H_1: \ \sigma_\alpha^2 \neq 0,$$

parece que no hay razón f apropiada, a menos que se encontrara que uno o más de los componentes de la varianza de interacción de 2 factores no es significativo. Por ejemplo, suponga que se hubiera comparado s_5^2 (cuadrado medio AC) con s_7^2 (cuadrado medio ABC) y se encontrara que $\sigma_{\alpha\gamma}^2$ es despreciable. Entonces podría argumentarse que el término $\sigma_{\alpha\gamma}^2$ debería eliminarse de todos los cuadrados medios esperados de la tabla 14.15; entonces, la razón s_1^2/s_4^2 ofrece una prueba de la significancia del componente σ_α^2 de la varianza. Por lo tanto, si se prueba la hipótesis concerniente a los componentes de la varianza de los efectos principales, es necesario investigar primero la significancia de los componentes de la interacción de 2 factores. Cuando se encuentra que ciertos componentes de la varianza de la interacción de 2 factores son significativos, por lo que deben permanecer como parte del cuadrado medio esperado, se utiliza una prueba aproximada derivada por Satterthwaite (1946; véase la bibliografía) .

Ejemplo 14.6: En un estudio realizado para determinar cuáles son las fuentes importantes de la variación en un proceso industrial, se toman 3 mediciones del producto para 3 operadores elegidos al azar, y se eligen en forma aleatoria 4 lotes de materia prima. Se decidió que debe hacerse una prueba estadística a un nivel de significancia de 0.05 para determinar si los componentes de la varianza debidos a los lotes, los operadores y la interacción son significativos. Además, tienen que calcularse los estimados de los componentes de la varianza. En la tabla 14.16 se presentan los datos con la respuesta expresada en porcentaje por peso:

Tabla 14.16: Datos para el ejemplo 14.6

	Lote			
Operador	**1**	**2**	**3**	**4**
1	66.9	68.3	69.0	69.3
	68.1	67.4	69.8	70.9
	67.2	67.7	67.5	71.4
2	66.3	68.1	69.7	69.4
	65.4	66.9	68.8	69.6
	65.8	67.6	69.2	70.0
3	65.6	66.0	67.1	67.9
	66.3	66.9	66.2	68.4
	65.2	67.3	67.4	68.7

***Solución*:** Las sumas de cuadrados se calculan de la forma acostumbrada y se obtienen los siguientes resultados:

STC (total) $= 84.5564$, SCE (error) $= 10.6733$,
SCA (operadores) $= 18.2106$, SCB (lotes) $= 50.1564$,
$SC(AB)$ (interacción) $= 5.5161$.

Se realizaron todos los demás cálculos y se presentan en la tabla 14.17. Como

$$f_{0.05}(2, 6) = 5.14, \qquad f_{0.05}(3, 6) = 4.76, \qquad \text{y} \qquad f_{0.05}(6, 24) = 2.51,$$

se descubre que los componentes de la varianza de los operadores y el lote son significativos. Aunque la varianza de la interacción no es significativa a un nivel $\alpha = 0.05$, el valor P es de 0.095. Los estimados de los componentes de la varianza del efecto principal son

$$\hat{\sigma}_{\alpha}^{2} = \frac{9.1053 - 0.9194}{12} = 0.68, \qquad \hat{\sigma}_{\beta}^{2} = \frac{16.7188 - 0.9194}{9} = 1.76.$$

Tabla 14.17: Análisis de varianza para el ejemplo 14.6

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Operadores	18.2106	2	9.1053	9.90
Lotes	50.1564	3	16.7188	18.18
Interacción	5.5161	6	0.9194	2.07
Error	10.6733	24	0.4447	
Total	84.5564	35		

Experimento del modelo mixto

Hay situaciones en que el experimento dicta la suposición de un **modelo mixto**, es decir, una mezcla de efectos aleatorios y fijos. Por ejemplo, para el caso de 2 factores se tiene que

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk},$$

para $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, n$. Las A_i pueden ser variables aleatorias independientes de ϵ_{ijk}, y las B_j pueden ser efectos fijos. La naturaleza mixta del modelo requiere que los términos de la interacción sean variables aleatorias. Como resultado, las hipótesis relevantes adoptan la forma

$$H_0': \sigma_{\alpha}^{2} = 0, \quad H_0'': B_1 = B_2 = \cdots = B_b = 0, \quad H_0''': \sigma_{\alpha\beta}^{2} = 0,$$
$$H_1': \sigma_{\alpha}^{2} \neq 0, \quad H_1'': \text{Al menos una de las } B_j \text{ no es igual a 0}, \quad H_1''': \sigma_{\alpha\beta}^{2} \neq 0.$$

Otra vez, los cálculos de la suma de cuadrados son idénticos a los de las situaciones de efectos fijos y aleatorios, y la prueba F es determinada por los cuadrados medios esperados. La tabla 14.18 proporciona los cuadrados medios esperados para el problema del modelo mixto de 2 factores.

Tabla 14.18: Cuadrados medios esperados para el experimento del modelo mixto de 2 factores

Factor	Cuadrado medio esperado
A (aleatorios)	$\sigma^2 + bn\sigma_{\alpha}^{2}$
B (fijos)	$\sigma^2 + n\sigma_{\alpha\beta}^{2} + \frac{an}{b-1}\sum_j B_j^2$
AB (aleatorios)	$\sigma^2 + n\sigma_{\alpha\beta}^{2}$
Error	σ^2

A partir de la naturaleza de los cuadrados medios esperados queda claro que la **prueba sobre el efecto aleatorio emplea el cuadrado medio del error s^2** como denominador, mientras que la **prueba sobre el efecto fijo** utiliza el cuadrado medio de interacción. Suponga que ahora se consideran 3 factores. En este caso, por supuesto, debe tomarse en cuenta la situación en que un factor es fijo y la situación en que 2 factores son fijos. La tabla 14.19 cubre ambas situaciones.

Tabla 14.19: Cuadrados medios esperados para experimentos factoriales de modelo mixto de 3 factores

	***A* aleatoria**	***A* aleatoria, *B* aleatoria**
A	$\sigma^2 + bcn\sigma^2_\alpha$	$\sigma^2 + cn\sigma^2_{\alpha\beta} + bcn\sigma^2_\alpha$
B	$\sigma^2 + cn\sigma^2_{\alpha\beta} + acn\sum_{j=1}^{b}\frac{B_j^2}{b-1}$	$\sigma^2 + cn\sigma^2_{\alpha\beta} + acn\sigma^2_\beta$
C	$\sigma^2 + bn\sigma^2_{\alpha\gamma} + abn\sum_{k=1}^{c}\frac{C_k^2}{c-1}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma} + bn\sigma^2_{\alpha\gamma} + abn\sum_{k=1}^{c}\frac{C_k^2}{c-1}$
AB	$\sigma^2 + cn\sigma^2_{\alpha\beta}$	$\sigma^2 + cn\sigma^2_{\alpha\beta}$
AC	$\sigma^2 + bn\sigma^2_{\alpha\gamma}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma}$
BC	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sum_j\sum_k\frac{(BC)_{jk}^2}{(b-1)(c-1)}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma}$
AB	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$
Error	σ^2	σ^2

Observe que en el caso de *A* aleatoria todos los efectos tienen pruebas *f* apropiadas. No obstante, para *A* y *B* aleatorias, el efecto principal *C* debe probarse utilizando un procedimiento tipo Satterthwaite, similar al que se emplea en el experimento de efectos aleatorios.

Ejercicios

14.26 Suponga un experimento de efectos aleatorios para el ejercicio 14.2 de la página 575 y estime los componentes de la varianza para las marcas de concentrado de jugo de naranja, para el número de días transcurridos a partir del día en que se mezcló el jugo hasta el día en que se hizo la prueba, y para el error experimental.

14.27 Para estimar los diversos componentes de la variabilidad en un proceso de filtración el porcentaje de material que se pierde en el licor madre se mide en 12 condiciones experimentales, con 3 corridas en cada condición. Se seleccionan al azar 3 filtros y 4 operadores para usarlos en el experimento.

a) Pruebe la hipótesis de que no hay un componente de interacción de la varianza entre los filtros y los operadores a un nivel de significancia $\alpha = 0.05$.

b) Pruebe la hipótesis de que los operadores y los filtros no tienen ningún efecto sobre la variabilidad del proceso de filtración a un nivel de significancia $\alpha = 0.05$.

c) Estime los componentes de la varianza que se deben a los filtros, a los operadores y al error experimental.

	Operador			
Filtro	**1**	**2**	**3**	**4**
1	16.2	15.9	15.6	14.9
	16.8	15.1	15.9	15.2
	17.1	14.5	16.1	14.9
2	16.6	16.0	16.1	15.4
	16.9	16.3	16.0	14.6
	16.8	16.5	17.2	15.9
3	16.7	16.5	16.4	16.1
	16.9	16.9	17.4	15.4
	17.1	16.8	16.9	15.6

14.28 Un contratista de la defensa está interesado en estudiar un proceso de inspección para detectar la falla o la fatiga de partes de recambio. Se utilizan 3 niveles de inspección que ejecutan 3 inspectores elegidos al azar. Se emplean 5 lotes para cada combinación en el estudio. Los niveles de los factores están en los datos.

La respuesta se expresa en fallas por cada 1000 piezas.

a) Escriba un modelo adecuado, con suposiciones.

b) Utilice análisis de varianza para probar las hipótesis apropiadas para los inspectores, el nivel de inspección y la interacción.

	Nivel de inspección		
Inspector	**Inspección militar completa**	**Inspección militar reducida**	**Comercial**
A	7.50 7.42	7.08 6.17	6.15 5.52
	5.85 5.89	5.65 5.30	5.48 5.48
	5.35	5.02	5.98
B	7.58 6.52	7.68 5.86	6.17 6.20
	6.54 5.64	5.28 5.38	5.44 5.75
	5.12	4.87	5.68
C	7.70 6.82	7.19 6.19	6.21 5.66
	6.42 5.39	5.85 5.35	5.36 5.90
	5.35	5.01	6.12

14.29 Considere el análisis de varianza siguiente para un experimento de efectos aleatorios:

Fuente de variación	Grados de libertad	Cuadrado medio
A	3	140
B	1	480
C	2	325
AB	3	15
AC	6	24
BC	2	18
ABC	6	2
Error	24	5
Total	47	

Pruebe si existen componentes significativos de la varianza entre todos los efectos principales y los efectos de interacción a un nivel de significancia de 0.01,

a) utilice un estimado agrupado del error cuando esto sea apropiado;

b) sin agrupar las sumas de los cuadrados de efectos insignificantes.

14.30 A un gerente de una planta le gustaría demostrar que la producción de una fábrica de lana de su planta no depende del operador de la máquina ni de la hora del día, y que es consistentemente elevada. Se eligen al azar 4 operadores y 3 horas del día para el estudio. Se mide el producto en yardas por minuto y se toman muestras 3 días elegidos al azar.

a) Escriba el modelo apropiado.

b) Evalúe los componentes de la varianza para el operador y la hora.

c) Saque sus conclusiones.

	Operador			
Hora	**1**	**2**	**3**	**4**
1	9.5	9.8	9.8	10.0
	9.8	10.1	10.3	9.7
	10.0	9.6	9.7	10.2
2	10.2	10.1	10.2	10.3
	9.9	9.8	9.8	10.1
	9.5	9.7	9.7	9.9
3	10.5	10.4	9.9	10.0
	10.2	10.2	10.3	10.1
	9.3	9.8	10.2	9.7

14.31 Un fabricante de pintura de látex para interiores (marca *A*) quisiera demostrar que su pintura es más robusta para el material donde se aplica, que la de sus 2 competidores más cercanos. La respuesta es el tiempo, en años, hasta que comienza a picarse. El estudio incluye las 3 marcas de pintura y 3 materiales seleccionados al azar. Para cada combinación se utilizan 2 piezas.

	Marca de pintura		
Material	***A***	***B***	***C***
A	5.50 5.15	4.75 4.60	5.10 5.20
B	5.60 5.55	5.50 5.60	5.40 5.50
C	5.40 5.48	5.05 4.95	4.50 4.55

a) ¿Cómo se le llama a este tipo de modelo?

b) Analice los datos usando el modelo apropiado.

c) ¿Los datos apoyan la afirmación del fabricante de la marca *A*?

14.32 Un ingeniero de procesos desea determinar si el ajuste de potencia de las máquinas que se usan para llenar ciertos tipos de cajas de cereal tienen un efecto significativo sobre el peso real del producto. El estudio consta de 3 tipos de cereal elaborados por la empresa, elegidos al azar, y 3 flujos fijos de energía. Para cada combinación se mide el peso de 4 cajas de cereal diferentes seleccionadas al azar. El peso deseado es de 400 gramos. A continuación se presentan los datos.

Ajuste de potencia	Tipo del cereal		
	1	**2**	**3**
Bajo	395 390	392 392	402 405
	401 400	394 401	399 399
Actual	396 399	390 392	404 403
	400 402	395 502	400 399
Alto	410 408	404 406	415 412
	408 407	401 400	413 415

a) Proporcione el modelo adecuado y liste las suposiciones que se hacen.

b) ¿Hay un efecto significativo debido al ajuste de potencia?

c) ¿Existe un componente de la varianza significativo debido al tipo de cereal?

Ejercicios de repaso

14.33 El Centro de Consulta Estadística de Virginia Tech participó en el análisis de un conjunto de datos tomados por el personal del Departamento de Nutrición Humana y Alimentos, al cual le interesaba estudiar los efectos del tipo de harina y el porcentaje de edulcorante sobre ciertos atributos físicos de un tipo de pastel. Se usó harina multiusos y para pasteles, y el porcentaje de edulcorante varió en 4 niveles. Los siguientes datos presentan información acerca de la gravedad específica de las muestras de pastel. Se prepararon 3 pasteles con cada una de las 8 combinaciones de factores.

Concentración de edulcorante	Harina					
	Multiusos			Para pasteles		
0	0.90	0.87	0.90	0.91	0.90	0.80
50	0.86	0.89	0.91	0.88	0.82	0.83
75	0.93	0.88	0.87	0.86	0.85	0.80
100	0.79	0.82	0.80	0.86	0.85	0.85

a) Realice un análisis de varianza con 2 factores. Pruebe si existen diferencias entre los tipos de harina. Pruebe si hay diferencias entre las concentraciones de edulcorante.

b) Analice el efecto de la interacción, si lo hubiera. Proporcione valores *P* para todas las pruebas.

14.34 Se llevó a cabo un experimento en el Departamento de Ciencias de Alimentos de Virginia Tech. El objetivo fue caracterizar la textura de cierto tipo de pescado de la familia de los arenques. También se estudió el efecto de los tipos de salsa empleada para preparar el pescado. La respuesta en el experimento era un "valor de textura", medido con una máquina que rebanaba el producto de los peces. Los siguientes datos son los valores de textura:

Tipo de salsa	Tipo de pescado					
	Sábalo sin curar		Sábalo curado		Arenque	
Crema ácida	27.6	57.4	64.0	66.9	107.0	83.9
	47.8	71.1	66.5	66.8	110.4	93.4
	53.8		53.8		83.1	
Salsa envinada	49.8	31.0	48.3	62.2	88.0	95.2
	11.8	35.1	54.6	43.6	108.2	86.7
	16.1		41.8		105.2	

a) Haga un análisis de varianza. Determine si hay o no interacción entre el tipo de salsa y el tipo de pescado.

b) Con base en los resultados del inciso *a* y en pruebas *F* de los efectos principales, determine si hay una diferencia significativa en la textura debido a los tipos de salsa, y determine si existe una diferencia significativa entre los tipos de pescado.

14.35 Se hizo un estudio para determinar si las condiciones de humedad afectan la fuerza que se requiere para separar piezas de plástico engomadas. Se probaron 3 tipos de plástico con 4 niveles de humedad. Los resultados, en kilogramos, son los siguientes:

Tipo de plástico	Humedad			
	30%	50%	70%	90%
A	39.0	33.1	33.8	33.0
	42.8	37.8	30.7	32.9
B	36.9	27.2	29.7	28.5
	41.0	26.8	29.1	27.9
C	27.4	29.2	26.7	30.9
	30.3	29.9	32.0	31.5

a) Suponga un experimento de efectos fijos, realice un análisis de varianza y pruebe la hipótesis de que no hay interacción entre la humedad y el tipo de plástico a un nivel de significancia de 0.05.

b) Utilice sólo los plásticos *A* y *B* y el valor de s^2 del inciso *a* y vuelva a probar la presencia de interacción a un nivel de significancia de 0.05.

14.36 Personal del Departamento de Ingeniería de Materiales de Virginia Tech llevó a cabo un experimento para estudiar los efectos de los factores ambientales sobre la estabilidad de cierto tipo de aleación cobre-níquel. La respuesta básica fue la vida de fatiga del material. Los factores son el nivel de esfuerzo y el ambiente. Los datos son los siguientes:

Ambiente	Nivel de esfuerzo		
	Bajo	Medio	Alto
Hidrógeno seco	11.08	13.12	14.18
	10.98	13.04	14.90
	11.24	13.37	15.10
Humedad elevada (95%)	10.75	12.73	14.15
	10.52	12.87	14.42
	10.43	12.95	14.25

a) Haga un análisis de varianza para probar la interacción entre los factores. Use $\alpha = 0.05$.

b) Con base en el inciso *a*) efectúe un análisis sobre los 2 efectos principales y saque sus conclusiones. Utilice el método del valor *P* para sus conclusiones.

14.37 En el experimento del ejercicio de repaso 14.33 también se utilizó el volumen del pastel como respuesta. Las unidades en que se expresa son pulgadas cúbicas. Pruebe la interacción entre los factores y analice los efectos principales. Suponga que los 2 factores son efectos fijos

Concentración de edulcorante	Harina					
	Multiusos			Para pasteles		
0	4.48	3.98	4.42	4.12	4.92	5.10
50	3.68	5.04	3.72	5.00	4.26	4.34
75	3.92	3.82	4.06	4.82	4.34	4.40
100	3.26	3.80	3.40	4.32	4.18	4.30

14.38 Una válvula de control necesita ser muy sensible al voltaje de entrada para así generar un voltaje de salida adecuado. Un ingeniero gira las perillas de control para cambiar el voltaje de entrada. En el libro *SN-Ratio for the Quality Evaluation*, publicado por la Japanese Standards Association (1988), se describe un estudio sobre la forma en que esos 3 factores (posición relativa de las perillas de control, rango de control de las perillas y voltaje de entrada) influyen en la sensibilidad de una válvula de control. A continuación se presentan los factores y sus niveles. Los datos se refieren a la sensibilidad de una válvula de control.

Factor *A*: posición relativa de las perillas de control: centro −0.5, centro y centro + 0.5

Factor *B*: rango de control de las perillas: 2, 4.5 y 7 (mm)

Factor *C*: voltaje de entrada: 100, 120 y 150 (V)

		C					
A	*B*	C_1		C_2		C_3	
A_1	B_1	151	135	151	135	151	138
A_1	B_2	178	171	180	173	181	174
A_1	B_3	204	190	205	190	206	192
A_2	B_1	156	148	158	149	158	150
A_2	B_2	183	168	183	170	183	172
A_2	B_3	210	204	211	203	213	204
A_3	B_1	161	145	162	148	163	148
A_3	B_2	189	182	191	184	192	183
A_3	B_3	215	202	216	203	217	205

Realice un análisis de varianza con $\alpha = 0.05$ para probar la existencia de efectos principales y de interacción significativos. Saque sus conclusiones.

14.39 En el ejercicio 14.25 de la página 588 se describe un experimento que implica la extracción de polietileno a través de un solvente.

		Tiempo					
Solvente		**4**		**8**		**16**	
Etanol	**120**	94.0	94.0	93.8	94.2	91.1	90.5
	80	95.3	95.1	94.9	95.3	92.5	92.4
Tolueno	**120**	94.6	94.5	93.6	94.1	91.1	91.0
	80	95.4	95.4	95.6	96.0	92.1	92.1

a) Haga una clase diferente de análisis de los datos. Ajuste un modelo adecuado de regresión con una variable categórica del solvente, un término de temperatura, un término de tiempo, una interacción de la temperatura y el tiempo, una interacción del solvente y la temperatura y una interacción del solvente y el tiempo. Realice pruebas *t* para todos los coeficientes y describa sus hallazgos.

b) ¿Sus resultados sugieren que el etanol y el tolueno requieren modelos diferentes, o son equivalentes aparte de las intersecciones? Explique su respuesta.

c) ¿Encontró alguna conclusión que contradiga las conclusiones que sacó de la solución del ejercicio 14.25? Explique su respuesta.

14.40 En el libro *SN-Ratio for the Quality Evaluation*, publicado por la Japanese Standards Association (1988), se describe un estudio acerca de cómo la presión del aire de los neumáticos afecta la maniobrabilidad de un automóvil. Se compararon 3 presiones distintas de aire en los neumáticos sobre 3 superficies diferentes de manejo. Las 3 presiones del aire fueron: los neumáticos tanto del lado izquierdo como del derecho inflados a 6 kgf/cm^2, los neumáticos del lado izquierdo inflados a 6 kgf/cm^2 y los del lado derecho inflados a 3 kgf/cm^2, y los neumáticos de ambos lados inflados a 3 kgf/cm^2. Las tres superficies de manejo fueron asfalto, asfalto seco y cemento seco. Se observó 2 veces el radio de giro de un vehículo de prueba para cada nivel de presión de los neumáticos sobre cada una de las 3 superficies de manejo.

Superficie de manejo	**Presión del aire de los neumáticos**					
	1		**2**		**3**	
Asfalto	44.0	25.5	34.2	37.2	27.4	42.8
Asfalto seco	31.9	33.7	31.8	27.6	43.7	38.2
Cemento seco	27.3	39.5	46.6	28.1	35.5	34.6

Realice un análisis de varianza con los datos anteriores. Haga comentarios acerca de la interpretación de los efectos principales y de interacción.

14.41 El fabricante de cierta marca de café secado por congelación espera reducir el tiempo del proceso sin arriesgar la integridad del producto. El ingeniero de procesos desea usar 3 temperaturas para la cámara de secado y 4 tiempos de secado. El tiempo de secado actual es de 3 horas a una temperatura de −15°C. La respuesta del sabor es un promedio de las calificaciones de 4 jueces profesionales. La calificación está en una escala de 1 a 10, donde 10 es la mejor. En la tabla que sigue se presentan los datos.

	Temperatura					
Tiempo	**−20°C**		**−15°C**		**−10°C**	
1 hr	9.60	9.63	9.55	9.50	9.40	9.43
1.5 hr	9.75	9.73	9.60	9.61	9.55	9.48
2 hr	9.82	9.93	9.81	9.78	9.50	9.52
3 hr	9.78	9.81	9.80	9.75	9.55	9.58

a) ¿Qué tipo de modelo se debe utilizar? Plantee las suposiciones.

b) Analice los datos en forma apropiada.

c) Redacte un breve informe para el vicepresidente encargado y hágale una recomendación para la elaboración futura de este producto.

14.42 Para garantizar el número de cajeros necesarios durante las horas pico de operación, un banco urbano recabó datos. Se estudiaron 4 cajeros durante 3 horarios "ocupados", 1) entre semana, de 10:00 a 11:00 A.M., 2) por las tardes entre semana, entre las 2:00 y las 3:00 P.M., y 3) las mañanas de los sábados, entre 11:00 y las 12:00. Un analista eligió al azar 4 horarios

dentro de cada uno de los 3 periodos, para cada una de las 4 posiciones de los cajeros durante varios meses y se observó el número de clientes atendidos. Los datos son los siguientes:

	Periodo											
Cajero	**1**				**2**				**3**			
1	18	24	17	22	25	29	23	32	29	30	21	34
2	16	11	19	14	23	32	25	17	27	29	18	16
3	12	19	11	22	27	33	27	24	25	20	29	15
4	11	9	13	8	10	7	19	8	11	9	17	9

Se supone que el número de clientes atendidos es una variable aleatoria de Poisson.

a) Comente sobre el riesgo de llevar a cabo un análisis de varianza estándar con los datos anteriores. ¿Qué suposiciones, si las hubiera, se violarían?

b) Elabore una tabla de ANOVA estándar que incluya pruebas F de los efectos principales y las interacciones. Si las interacciones y los efectos principales resultan significativos, establezca las conclusiones científicas. ¿Qué aprendimos? Asegúrese de interpretar cualquier interacción significativa. Utilice su propio juicio respecto a los valores P.

c) Vuelva a hacer el análisis completo usando una transformación apropiada de la respuesta. ¿Encontró alguna diferencia en los resultados? Haga comentarios al respecto.

14.6 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Uno de los temas más susceptibles de confusión en el análisis de experimentos factoriales radica en la interpretación de los efectos principales ante la presencia de interacción. La existencia de un valor P relativamente grande para un efecto principal, cuando es clara la presencia de interacciones, podría tentar al analista a concluir que "no existe efecto principal significativo". Sin embargo, debe entenderse que si un efecto principal está implicado en una interacción significativa, entonces el efecto principal **está influyendo en la respuesta**. La naturaleza del efecto es inconsistente a través de los niveles de otros efectos. La naturaleza del papel que desempeña el efecto principal se deduce de las **gráficas de interacción**.

Debido a lo que se expresa en el párrafo anterior, hay un gran peligro de usar la estadística de manera equivocada cuando se emplea una prueba de comparación múltiple sobre los efectos principales ante la presencia clara de interacción entre los factores.

Debe tenerse precaución en el análisis de un experimento factorial cuando se supone un diseño completamente aleatorizado y en realidad no se hizo tal aleatorización. Por ejemplo, es común que se encuentren factores que son **muy difíciles de cambiar**. Como resultado, podría ser necesario mantener sin cambio los niveles de factores durante largos periodos a lo largo de todo el experimento. El ejemplo más común es el factor temperatura. Subirla o bajarla en un esquema aleatorio es un plan costoso y la mayoría de los experimentadores evitarán hacerlo. Los diseños experimentales con *restricciones en la aleatorización* son muy comunes y reciben el nombre de **diseños de gráficas separadas**. Esos diseños rebasan el alcance de este libro, pero en Montgomery (2008a) se encuentra su presentación.

Muchos de los conceptos que se analizaron en este capítulo se utilizarán en el capítulo 15, por ejemplo, la importancia de la aleatorización y el papel que desempeña la interacción en la interpretación de los resultados. Sin embargo, en el capítulo 15 se cubren 2 áreas que representan una expansión de los principios que se estudiaron en este capítulo y en el capítulo 13. En el capítulo 15 la solución de problemas con el uso de experimentos factoriales se realiza por medio del análisis de regresión, ya que se supone que la mayoría de los factores son cuantitativos y que se miden en un continuo, como la temperatura y el tiempo. Se derivan ecuaciones de predicción a partir de los datos del experimento diseñado y se utilizan para la mejora de procesos o incluso para su optimización. Además, se estudia el tema de los factoriales fraccionarios, en los que sólo una parte o fracción de todo el experimento factorial se aplica debido al costo excesivo que implica la realización de todo el experimento.

Capítulo 15

Experimentos factoriales 2^k y fracciones

15.1 Introducción

Ya se han expuesto ciertos conceptos del diseño experimental. El plan de muestreo para la prueba t simple sobre la media de una población normal y el análisis de varianza implican la asignación aleatoria de los tratamientos preseleccionados a las unidades experimentales. El diseño de bloques aleatorizados, en el que los tratamientos se asignan a las unidades dentro de bloques relativamente homogéneos implica una aleatorización restringida.

En este capítulo se presta atención especial a los diseños experimentales en los que el plan experimental requiere estudiar el efecto sobre una respuesta de k factores, cada uno en dos niveles. A éstos se les conoce como **experimentos factoriales 2^k**. Es frecuente que los niveles se denoten por "alto" y "bajo", aunque esa notación podría ser arbitraria en el caso de variables cualitativas. El diseño factorial completo requiere que cada nivel de cada factor ocurra con cada nivel de cada uno de los demás factores, lo que da un total de **2^k combinaciones de tratamientos**.

Filtrado de factores y experimentación secuencial

A menudo, cuando se realizan experimentos, ya sea en una investigación o a un nivel de desarrollo, un diseño experimental bien planeado corresponde a una **etapa** de lo que en realidad es el **plan secuencial** de la experimentación. Lo más frecuente al comienzo de un estudio es que los científicos e ingenieros no estén conscientes de cuáles factores son importantes ni de cuáles son los rangos apropiados para los factores potenciales sobre los que deben realizar la experimentación. Por ejemplo, en el libro *Response Surface Methodology*, Myers, Montgomery y Anderson-Cook (2009) dan un ejemplo de una investigación realizada en una planta piloto, la cual incluye un experimento en el que cuatro factores, temperatura, presión, concentración de formaldehído y tasa de agitación, se varían para establecer su influencia sobre la respuesta, es decir, la tasa de filtración de cierto producto químico. Incluso al nivel de planta piloto los científicos no están seguros respecto a si deben incluir los 4 factores en el modelo. Además, el objetivo final consiste en determinar la configuración adecuada de los factores contribuyentes que maximice la tasa de filtración. Por lo tanto, es necesario determinar

la **región apropiada de experimentación**. Estas preguntas sólo pueden responderse si todo el plan experimental se realiza en forma secuencial. Muchos procesos experimentales son planes que implican un *aprendizaje iterativo*, el tipo de aprendizaje consistente con el método científico, en el que la palabra *iterativo* implica experimentación por etapas.

Por lo común la primera etapa del plan secuencial ideal es variable o de **filtrado de factores**, un procedimiento que implica un diseño experimental de bajo costo en el que se utilizan **factores candidatos**. Esto es especialmente importante cuando el plan requiere un sistema complejo, como un proceso de manufactura. La información obtenida a partir de los resultados de un *diseño de filtrado* se emplea para diseñar uno o más experimentos posteriores, en los que se realizan ajustes de los factores importantes, los cuales proporcionan mejorías en el sistema o en el proceso.

Los experimentos factoriales 2^k y fracciones de 2^k son poderosas herramientas que constituyen diseños de filtrado ideales; son sencillos y prácticos, y atraen por intuición. Muchos de los conceptos generales que se estudian en el capítulo 14 siguen siendo válidos. Sin embargo, hay métodos gráficos que brindan información útil para el análisis de los diseños de 2 niveles.

Diseños de selección para cantidades grandes de factores

Cuando k es pequeña, digamos $k = 2$ o incluso $k = 3$, es evidente la utilidad del factorial 2^k para el filtrado de factores. Tanto el análisis de varianza como el de regresión, que se estudiaron e ilustraron en los capítulos 12, 13 y 14, continúan siendo herramientas útiles. Además, los enfoques gráficos también pueden ser de ayuda.

Si k es grande, por ejemplo 6, 7 u 8, el número de combinaciones de factores y, por lo tanto, de corridas experimentales necesarias para el factorial 2^k con frecuencia se vuelve prohibitivo. Por ejemplo, suponga que hay interés en realizar un diseño de selección que involucre $k = 8$ factores. Podría desearse obtener información acerca de todos los $k = 8$ efectos principales, así como de las $\frac{k(k-1)}{2} = 28$ interacciones de dos factores. Sin embargo, incluso $2^8 = 256$ corridas parecería que hace al estudio demasiado grande y excesivo para estudiar $28 + 8 = 36$ efectos. No obstante, como se verá en secciones posteriores, cuando k es grande es posible obtener gran cantidad de información de manera eficaz usando sólo una fracción del experimento factorial 2^k completo. Esta clase de diseños constituye la clase de *diseños factoriales fraccionarios*. La meta consiste en recuperar información de alta calidad acerca de los efectos principales y las interacciones interesantes, aun cuando el tamaño del diseño se reduzca en forma considerable.

15.2 El factorial 2^k: cálculo de efectos y análisis de varianza

Considere inicialmente un factorial 2^2 con factores A y B, y *n observaciones experimentales por combinación de factores*. Es útil emplear los símbolos (1), a, b y ab para denotar los puntos del diseño, donde la presencia de una letra minúscula implica que el factor (A o B) está en el *nivel alto*. Así, la ausencia de la minúscula implica que el factor está en el *nivel bajo*. Por lo que ab es el punto de diseño $(+, +)$, a es $(+, -)$, b es $(-, +)$ y (1) es $(-, -)$. Asimismo existen situaciones en las que la notación también se aplica

para los datos de respuesta en el punto de diseño en cuestión. Como introducción al cálculo de **efectos** importantes que ayuden a determinar la influencia de los factores y **sumas de cuadrados** que están incorporados en los cálculos del análisis de varianza se presenta la tabla 15.1.

Tabla 15.1: Un experimento factorial 2^2

		A		**Media**
B		b	ab	$\frac{b+ab}{2n}$
		(1)	a	$\frac{(1)+a}{2n}$
Media		$\frac{(1)+b}{2n}$	$\frac{a+ab}{2n}$	

En esta tabla, (1), a, b y ab representan totales de los n valores de la respuesta en los puntos de diseño individuales. La simplicidad del factorial 2^2 reside en el hecho de que, aparte del error experimental, el analista obtiene la información importante en componentes con un solo grado de libertad, uno para cada uno de los dos efectos principales A y B, y un grado de libertad para la interacción AB. La información que se recupera sobre todos estos aspectos adopta la forma de tres **contrastes**. Se definirán los siguientes contrastes entre los totales de los tratamientos:

$$\begin{aligned} \text{contraste } A &= ab + a - b - (1), \\ \text{contraste } B &= ab - a + b - (1), \\ \text{contraste } AB &= ab - a - b + (1). \end{aligned}$$

Los tres **efectos** del experimento implican estos contrastes y apelan al sentido común y a la intuición. Los dos efectos principales calculados tienen la forma

$$\text{efecto} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ y $\bar{y}_L$ son las respuestas promedio en el nivel alto o "+" y en el nivel bajo o "−", respectivamente. Como resultado,

Cálculo de los efectos principales

$$A = \frac{ab + a - b - (1)}{2n} = \frac{\text{contraste } A}{2n}$$

y

$$B = \frac{ab - a + b - (1)}{2n} = \frac{\text{contraste } B}{2n}.$$

La cantidad A es considerada la *diferencia entre la respuesta media en los niveles alto y bajo del factor A*. De hecho, A se denomina **efecto principal** del factor A. En forma similar, B es el efecto principal del factor B. Al inspeccionar la diferencia entre $ab - b$ y $a - (1)$ o entre $ab - a$ y $b - (1)$ en la tabla 15.1, se observa una aparente interacción en los datos. Si, por ejemplo,

$$ab - a \approx b - (1) \qquad \text{o bien} \qquad ab - a - b + (1) \approx 0,$$

una recta que conecta las respuestas para cada nivel del factor A en el nivel alto del factor B será aproximadamente paralela a una recta que conecte la respuesta para cada nivel del factor A en el nivel bajo del factor B. Las rectas no paralelas de la figura 15.1 sugieren la presencia de interacción. Para probar que esta interacción aparente es significativa se construye un tercer contraste en los totales del tratamiento, ortogonal a los contrastes del efecto principal, al cual se denomina **efecto de interacción**. La construcción del tercer contraste mencionado se realiza evaluando

Efecto de interacción

$$AB = \frac{ab - a - b + (1)}{2n} = \frac{\text{contraste } AB}{2n}.$$

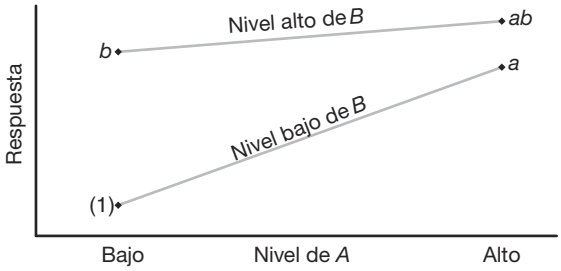


Figura 15.1: Respuesta que sugiere una interacción aparente.

Ejemplo 15.1: Considere los datos de las tablas 15.2 y 15.3 con $n = 1$ para un experimento factorial 2^2.

Tabla 15.2: Factorial 2^2 sin interacción

	B	
A	−	+
+	50	70
−	80	100

Tabla 15.3: Factorial 2^2 con interacción

	B	
A	−	+
+	50	70
−	80	40

Los números en las celdas de las tablas 15.2 y 15.3 ilustran con claridad la manera en que los contrastes y el cálculo resultante de los dos efectos principales y de las conclusiones resultantes pueden estar muy influidos por la presencia de interacción. En la tabla 15.2 el efecto de A es -30 tanto en el nivel bajo como en el nivel alto del factor B, y el efecto de B es 20 en los niveles bajo y alto del factor A. Esta "consistencia del efecto" (no hay interacción) puede ser información muy importante para el analista. Los efectos principales son

$$A = \frac{70 + 50}{2} - \frac{100 + 80}{2} = 60 - 90 = -30,$$
$$B = \frac{100 + 70}{2} - \frac{80 + 50}{2} = 85 - 65 = 20,$$

mientras que el efecto de la interacción es

$$AB = \frac{100 + 50}{2} - \frac{80 + 70}{2} = 75 - 75 = 0.$$

Por otro lado, en la tabla 15.3 el efecto A es nuevamente -30 al nivel bajo de B, pero $+30$ al nivel alto de B. Esta "inconsistencia del efecto" (interacción) también está presente para B en todos los niveles de A. En estos casos los efectos principales podrían carecer de significado y, de hecho, prestarse mucho a la confusión. Por ejemplo, el efecto de A es

$$A = \frac{50 + 70}{2} - \frac{80 + 40}{2} = 0,$$

ya que hay un "enmascaramiento" completo del efecto conforme se promedia sobre los niveles de B. La fuerte interacción se ilustra con el efecto calculado

$$AB = \frac{70 + 80}{2} - \frac{50 + 40}{2} = 30.$$

Aquí es conveniente ilustrar los escenarios de las tablas 15.2 y 15.3 con las gráficas de interacción. Observe el paralelismo en la gráfica de la figura 15.2 y la interacción aparente en la figura 15.3.

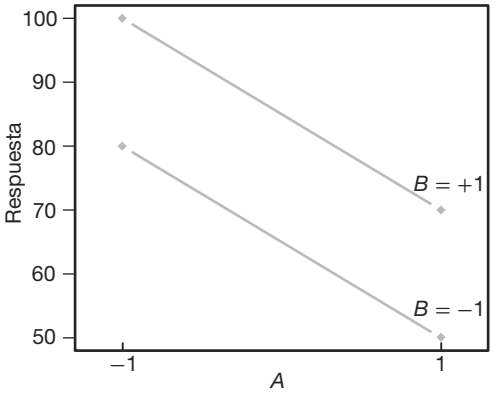


Figura 15.2: Gráfica de interacción para los datos de la tabla 15.2.

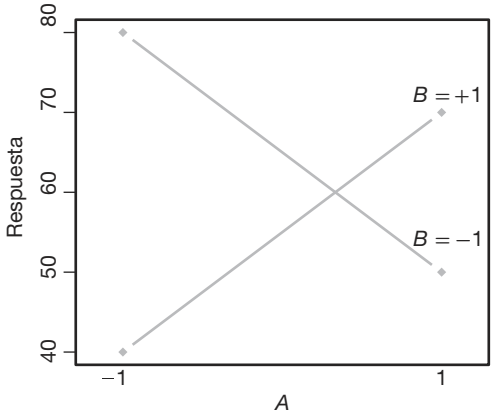


Figura 15.3: Gráfica de interacción para los datos de la tabla 15.3.

Cálculo de las sumas de cuadrados

Se aprovecha el hecho de que en el factorial 2^2, o para el caso en el experimento factorial 2^k general, cada efecto principal y efecto de interacción tiene asociado **un solo grado de libertad**. Por lo tanto, es posible escribir contrastes ortogonales $2^k - 1$ con un solo grado de libertad en las combinaciones de tratamientos, donde cada uno es responsable de la variación debida a cierto efecto principal o interacción. Así, con base en las suposiciones usuales de independencia y normalidad en el modelo experimental, se hacen pruebas para determinar si el contraste refleja variación sistemática, o bien, sólo variaciones probabilísticas o aleatorias. Las sumas de cuadrados para cada contraste se calculan siguiendo los procedimientos que se estudiaron en la sección 13.5. Si se escribe

$$Y_{1..} = b + (1), \qquad Y_{2..} = ab + a, \qquad c_1 = -1 \qquad \text{y} \qquad c_2 = 1,$$

donde $Y_{1..}$ y $Y_{2..}$ constituyen el total de $2n$ observaciones, se tiene

$$SCA = SCw_A = \frac{\left(\sum_{i=1}^{2} c_i Y_{i..}\right)^2}{2n\sum_{i=1}^{2} c_i^2} = \frac{[ab + a - b - (1)]^2}{2^2 n} = \frac{(\text{contraste } A)^2}{2^2 n},$$

con 1 grado de libertad. De forma similar, se encuentra que

$$SCB = \frac{[ab + b - a - (1)]^2}{2^2 n} = \frac{(\text{contraste } B)^2}{2^2 n}$$

y

$$SC(AB) = \frac{[ab + (1) - a - b]^2}{2^2 n} = \frac{(\text{contraste } AB)^2}{2^2 n}.$$

Cada contraste tiene 1 grado de libertad, mientras que las sumas de cuadrados del error, con $2^2(n - 1)$ grados de libertad, se obtienen mediante una resta a partir de la fórmula

$$SCE = SCT - SCA - SCB - SC(AB).$$

Al calcular las sumas de cuadrados para los efectos principales A y B, y el efecto de interacción AB, es conveniente presentar las salidas totales de las combinaciones de tratamiento junto con los signos algebraicos apropiados para cada contraste, como se observa en la tabla 15.4. Los efectos principales se obtienen como comparaciones simples entre los niveles alto y bajo. Por lo tanto, se asigna un signo positivo para la combinación de tratamientos que esté en el nivel alto de un factor dado, y uno negativo a la combinación de tratamientos del nivel bajo. Los signos positivo y negativo para el efecto de interacción se obtienen multiplicando los signos correspondientes de los contrastes de los factores de la interacción.

Tabla 15.4: Signos para los contrastes en un experimento factorial 2^2

Combinación de tratamientos	Efecto factorial		
	A	***B***	***AB***
(1)	−	−	+
a	+	−	−
b	−	+	−
ab	+	+	+

El factorial 2^3

Ahora consideremos un experimento en el que intervienen tres factores, A, B y C, cada uno con niveles -1 y $+1$. Se trata de un experimento factorial 2^3 que proporciona ocho combinaciones de tratamientos (1), a, b, c, ab, ac, bc y abc. En la tabla 15.5 se presentan las combinaciones de tratamientos y los signos algebraicos apropiados para cada contraste que se usan en el cálculo de las sumas de los cuadrados para los efectos principales y los efectos de interacción.

Tabla15.5:Signos de los contrastes en un experimento factorial 2^3

Combinación de tratamiento	Efecto factorial (simbólico)						
	A	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
(1)	−	−	−	+	+	+	−
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
ab	+	+	−	+	−	−	−
ac	+	−	+	−	+	−	−
bc	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+

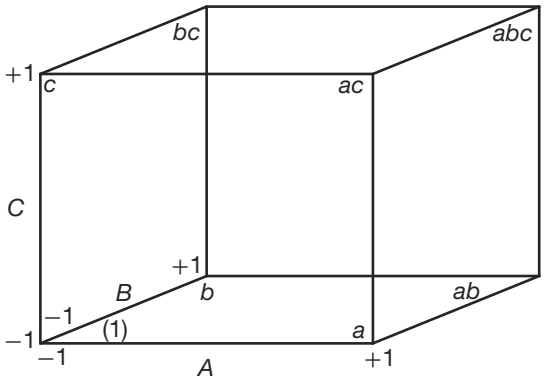


Figura 15.4: Vista geométrica de 2^3.

Es útil analizar e ilustrar la geometría del factorial 2^3 del mismo modo que se hizo para el factorial 2^2 en la figura 15.1. Para el 2^3 los **ocho puntos de diseño** representan los vértices de un cubo, como se observa en la figura 15.4.

Las columnas de la tabla 15.5 representan los signos que se utilizan para los contrastes, así como los cálculos de siete efectos y las sumas de cuadrados correspondientes. Estas columnas son análogas a las que se observan en la tabla 15.4 para el caso de 2^2. Como son ocho puntos de diseño hay siete efectos disponibles. Por ejemplo,

$$A = \frac{a + ab + ac + abc - (1) - b - c - bc}{4n},$$

$$AB = \frac{(1) + c + ab + abc - a - b - ac - bc}{4n},$$

y así sucesivamente. Las sumas de cuadrados son dadas por

$$SC(\text{efecto}) = \frac{(\text{contraste})^2}{2^3 n}.$$

Al observar la tabla 15.5 se revela que para el experimento 2^3 todos los contrastes

entre los siete son mutuamente ortogonales y, por lo tanto, los siete efectos se evalúan en forma independiente.

Efectos y sumas de cuadrados para el 2^k

Para un experimento factorial 2^k las sumas de cuadrados de un solo grado de libertad para los efectos principales y los efectos de interacción se obtienen elevando al cuadrado los contrastes apropiados en los totales del tratamiento y dividiendo entre $2^k n$, donde n es el número de réplicas de las combinaciones del tratamiento.

Como antes, un efecto siempre se calcula restando la respuesta promedio en el nivel "bajo", de la respuesta promedio en el nivel "alto". Quedan muy claros los niveles alto y bajo para los efectos principales. Los niveles alto y bajo simbólicos para las interacciones son evidentes a partir de la información de la tabla 15.5.

La propiedad de ortogonalidad tiene la misma importancia aquí que en el material sobre las comparaciones que se estudió en el capítulo 13. La ortogonalidad de los contrastes implica que los efectos estimados y, por lo tanto, las sumas de cuadrados, son independientes. Esta independencia se ilustra con claridad en el experimento factorial 2^3 si las respuestas, con el factor A en su nivel alto, se incrementan en una cantidad x en la tabla 15.5. Sólo el contraste A conduce a una suma de cuadrados más grande, ya que el efecto x se cancela cuando se forman los seis contrastes restantes como resultado de los dos signos positivos y los dos negativos asociados con las combinaciones de tratamientos en los que A está en el nivel alto.

La ortogonalidad produce otras ventajas, las cuales se abordarán cuando se estudie el experimento factorial 2^k en situaciones de regresión.

15.3 Experimento factorial 2^k sin réplicas

El factorial completo 2^k con frecuencia requiere mucha experimentación, en particular cuando k es grande. Como resultado, a menudo no es posible replicar cada combinación de factores. Si en el modelo del experimento se incluyen todos los efectos, con todas las interacciones, no se permite ningún grado de libertad para el error. A menudo, cuando k es grande, el analista de datos *agrupará* las sumas de los cuadrados y los grados de libertad correspondientes para las interacciones de orden superior que se sabe, o se supone, son despreciables. Esto producirá pruebas F para los efectos principales e interacciones de orden inferior.

Graficación de diagnóstico con experimentos factoriales 2^k sin réplicas

Las gráficas de probabilidad normal constituyen una metodología muy útil para determinar la importancia relativa de los efectos en un experimento con factores de dos niveles razonablemente grandes cuando no hay réplica. Este tipo de gráfica de diagnóstico puede ser útil sobre todo cuando el analista de datos duda en agrupar interacciones de orden superior por temor de agrupar en el "error" algunos efectos verdaderamente reales y no sólo aleatorios. El lector debe recordar que todos los efectos que no son reales, es decir, que son *estimados de cero* independientes, siguen una distribución normal con media cercana a cero y varianza constante. Por ejemplo, en un experimento factorial 2^4 se debe recordar que todos los efectos, teniendo en cuenta que $n = 1$, son de la forma

$$AB = \frac{\text{contraste}}{8} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ es el promedio de ocho corridas experimentales independientes en el nivel alto, o "+", y $\bar{y}_L$ es el promedio de ocho corridas independientes en el nivel bajo, o "−". Así, la varianza de cada contraste es $\text{Var}(\bar{y}_H - \bar{y}_L) = \sigma^2/4$. Para cualesquiera efectos reales $E(\bar{y}_H - \bar{y}_L) \neq 0$. Así, la gráfica de probabilidad normal debería revelar efectos "significativos" como aquellos que caen fuera de la línea recta que describe realizaciones de variables aleatorias normales independientes distribuidas de forma idéntica.

La gráfica de probabilidad puede adoptar una de muchas formas. Se recomienda al lector que consulte el capítulo 8, en el que se presentaron dichas gráficas por primera vez. Se puede usar la gráfica cuantil-cuantil, normal y empírica. También es posible utilizar el procedimiento de graficación que emplea el papel de probabilidad normal. Además, existen otros tipos de gráficas de probabilidad normal para el diagnóstico. En resumen, las gráficas de efectos para el diagnóstico son como sigue.

Gráficas de efectos de probabilidad para experimentos factoriales 2^4 sin réplica

1. Calcular los efectos como
$$\text{efecto} = \frac{\text{contraste}}{2^{k-1}}.$$
2. Construir una gráfica de probabilidad normal de todos los efectos.
3. Los efectos que caigan fuera de la línea recta deben considerarse reales.

A continuación se hacen más comentarios respecto de las gráficas de probabilidad normal de los efectos. En primer lugar, el analista podría sentirse frustrado si utiliza las gráficas con un experimento pequeño. Por otro lado, la graficación puede proporcionar resultados satisfactorios cuando hay *dispersión de efectos,* muchos efectos que no son verdaderamente reales. Esta dispersión será evidente en experimentos grandes, en los que es poco probable que las interacciones de orden superior sean reales.

Estudio de caso 15.1: **Moldeado por inyección.** Muchas empresas fabricantes de Estados Unidos y otros países utilizan partes moldeadas como componentes de un proceso. Un problema grande que enfrentan con frecuencia es el rebasamiento. A menudo, un molde troquelado de una parte se construye con un tamaño más grande que el nominal para permitir que se contraiga. En la siguiente situación experimental se produce un molde nuevo para el cual es importante encontrar las especificaciones adecuadas del proceso para minimizar la contracción. En el siguiente experimento los valores de la respuesta son desviaciones de los nominales, es decir, contracciones. Los factores y niveles son los siguientes:

	Niveles codificados	
	−1	**+1**
***A*. Velocidad de inyección (pies/seg)**	1.0	2.0
***B*. Temperatura de moldeado (°C)**	100	150
***C*. Presión de moldeado (psi)**	500	1000
***D*. Contrapresión (psi)**	75	120

El propósito del experimento fue determinar cuáles efectos (principales y de interacción) influyen en la contracción. El experimento se consideró un filtrado preliminar a partir del cual se determinaron los factores para un análisis más completo. Asimismo, se espera obtener información respecto a cómo los factores importantes repercuten en la contracción. En la tabla 15.6 se presentan los datos de un experimento factorial 2^4 sin réplica.

Tabla 15.6: Datos para el estudio de caso 15.1

Combinación de factores	Respuesta (cm $\times$ 10^4)	Combinación de factores	Respuesta (cm $\times$ 10^4)
(1)	72.68	*d*	73.52
a	71.74	*ad*	75.97
b	76.09	*bd*	74.28
ab	93.19	*abd*	92.87
c	71.25	*cd*	79.34
ac	70.59	*acd*	75.12
bc	70.92	*bcd*	79.67
abc	104.96	*abcd*	97.80

Inicialmente se calcularon los efectos y se plasmaron en una gráfica de probabilidad normal. Los efectos calculados son los siguientes:

$$\begin{aligned} A &= 10.5613, & BD &= -2.2787, & B &= 12.4463, \\ C &= 2.4138, & D &= 2.1438, & AB &= 11.4038, \\ AC &= 1.2613, & AD &= -1.8238, & BC &= 1.8163, \\ CD &= 1.4088, & ABC &= 2.8588, & ABD &= -1.7813, \\ ACD &= -3.0438, & BCD &= -0.4788, & ABCD &= -1.3063. \end{aligned}$$

En la figura 15.5 se observa la gráfica cuantil-cuantil normal, la cual parece implicar que los efectos *A*, *B* y *AB* son importantes. Los signos de los efectos importantes indican las conclusiones preliminares.

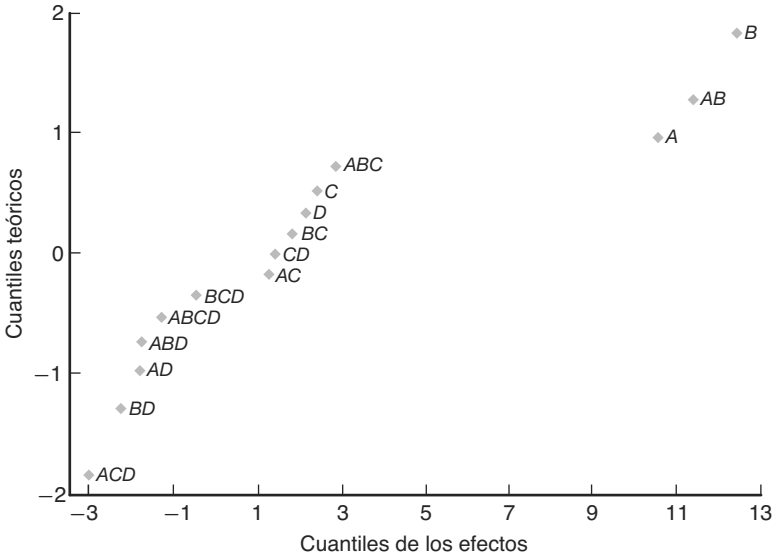


Figura 15.5: Gráfica cuantil-cuantil normal de los efectos para el estudio de caso del ejemplo 15.1.

1. Un incremento en la velocidad de inyección de 1.0 a 2.0 aumenta la contracción.
2. Un aumento en la temperatura de moldeado de 100°C a 150°C incrementa la contracción.
3. Hay una interacción entre la velocidad de inyección y la temperatura del moldeado; aunque ambos efectos principales son importantes es crucial entender el efecto de la interacción de los dos factores.

Interpretación de la interacción de dos factores

Como se esperaría, una tabla de medias de dos factores facilita la interpretación de la interacción *AB*. Considere la situación de dos factores de la tabla 15.7.

Tabla 15.7: Ilustración de una interacción de dos factores

	***B* (temperatura)**	
***A* (velocidad)**	**100**	**150**
2	73.355	97.205
1	74.1975	75.240

Observe que la media muestral grande a velocidad y temperatura elevadas creó la interacción significativa. La **contracción se incrementa en forma no aditiva**. La temperatura del moldeado parece tener un efecto positivo a pesar del nivel de velocidad. Sin embargo, el efecto es el mayor a velocidad elevada. El efecto de la velocidad es muy ligero a temperaturas bajas, pero es claramente positivo a una temperatura elevada de moldeado. Para controlar la contracción a bajo nivel *debería evitarse el uso simultáneo de una alta velocidad de inyección y una temperatura de moldeado elevada*. Todos estos resultados se ilustran en forma gráfica en la figura 15.6.

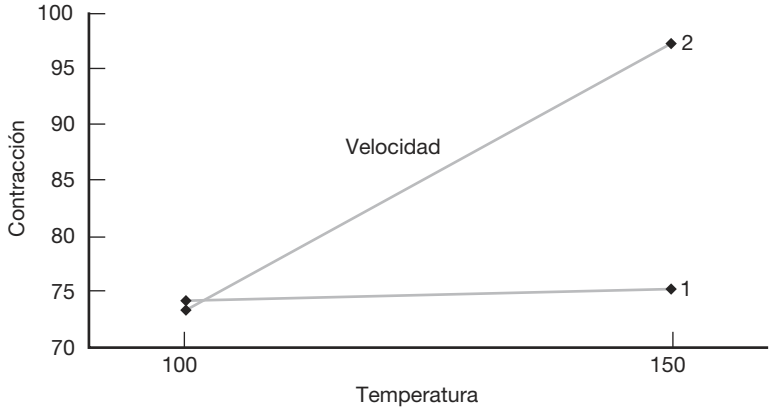


Figura 15.6: Gráfica de la interacción para el estudio de caso 15.1.

Análisis con el cuadrado medio del error agrupado: salida de resultados por computadora comentada

Puede ser de interés observar un análisis de varianza de los datos del moldeado por inyección con interacciones de orden superior agrupadas para formar un cuadrado medio del error. Las interacciones de órdenes tres y cuatro están agrupadas. En la figura 15.7 se observa una salida de resultados por computadora de la función PROC GLM del *SAS*. El análisis de varianza revela, en esencia, la misma conclusión que la gráfica de probabilidad normal.

Las pruebas y los valores *P* que se observan en la figura 15.7 requieren una interpretación. Un valor *P* significativo sugiere que el efecto difiere de cero en forma significativa. Las pruebas sobre los efectos principales (que en presencia de las interacciones se pueden considerar como los efectos promediados sobre los niveles de los demás factores) indican la significancia de los efectos *A* y *B*. Los signos de los efectos también son importantes. Un aumento en el nivel de bajo a alto en *A*, la velocidad de inyección, ocasiona

```
                        The GLM Procedure
Dependent Variable: y
                              Sum of
Source                DF     Squares  Mean Square  F Value  Pr > F
Model                 10  1689.237462   168.923746     9.37  0.0117
Error                  5    90.180831    18.036166
Corrected Total       15  1779.418294
R-Square     Coeff Var    Root MSE      y Mean
0.949320      5.308667    4.246901    79.99938
Source     DF   Type III SS   Mean Square  F Value    Pr > F
A           1   446.1600062   446.1600062    24.74    0.0042
B           1   619.6365563   619.6365563    34.36    0.0020
C           1    23.3047563    23.3047563     1.29    0.3072
D           1    18.3826563    18.3826563     1.02    0.3590
A*B         1   520.1820562   520.1820562    28.84    0.0030
A*C         1     6.3630063     6.3630063     0.35    0.5784
A*D         1    13.3042562    13.3042562     0.74    0.4297
B*C         1    13.1950562    13.1950562     0.73    0.4314
B*D         1    20.7708062    20.7708062     1.15    0.3322
C*D         1     7.9383063     7.9383063     0.44    0.5364
                             Standard
Parameter      Estimate         Error  t Value    Pr > |t|
Intercept   79.99937500    1.06172520    75.35 <.0001
A            5.28062500    1.06172520     4.97   0.0042
B            6.22312500    1.06172520     5.86   0.0020
C            1.20687500    1.06172520     1.14   0.3072
D            1.07187500    1.06172520     1.01   0.3590
A*B          5.70187500    1.06172520     5.37   0.0030
A*C          0.63062500    1.06172520     0.59   0.5784
A*D         -0.91187500    1.06172520    -0.86   0.4297
B*C          0.90812500    1.06172520     0.86   0.4314
B*D         -1.13937500    1.06172520    -1.07   0.3322
C*D          0.70437500    1.06172520     0.66   0.5364
```

Figura 15.7: Salida de resultados por computadora del *SAS* para los datos del estudio de caso 15.1.

un incremento en la contracción. Lo mismo es verdad para *B*. Sin embargo, debido a la interacción significativa *AB*, las interpretaciones del efecto principal podrían considerarse como tendencias en todos los niveles de los demás factores. El impacto de la interacción *AB* significativa se entiende mejor si se emplea una tabla de medias de dos factores.

Ejercicios

15.1 Los siguientes datos se obtuvieron de un experimento factorial 2^3 que se replicó tres veces. Utilice el método del contraste para evaluar las sumas de cuadrados de todos los efectos factoriales. Saque sus conclusiones.

Combinación de tratamientos	Réplica 1	Réplica 2	Réplica 3
(1)	12	19	10
a	15	20	16
b	24	16	17
ab	23	17	27
c	17	25	21
ac	16	19	19
bc	24	23	29
abc	28	25	20

15.2 En un experimento efectuado por el Departamento de Ingeniería de Minas de Virginia Tech con el fin de estudiar un sistema de filtrado particular para carbón se agregó un coagulante a la solución contenida en un tanque con carbón y sedimentos, que luego se puso en un sistema de recirculación para purificar el carbón. En el proceso experimental se variaron tres factores:

Factor *A*: porcentaje de sólidos que circularon inicialmente en el sobreflujo
Factor B: tasa de flujo del polímero
Factor C: pH del tanque

La cantidad de sólidos en el flujo inferior del sistema de purificación determina qué tan puro ha quedado el carbón. Se emplearon dos niveles de cada factor y se hicieron dos corridas experimentales para cada una de las $2^3 = 8$ combinaciones. En la siguiente tabla se especifican las mediciones de respuesta en porcentajes de sólidos por peso:

Combinación de tratamientos	Respuesta	
	Réplica 1	Réplica 2
(1)	4.65	5.81
a	21.42	21.35
b	12.66	12.56
ab	18.27	16.62
c	7.93	7.88
ac	13.18	12.87
bc	6.51	6.26
abc	18.23	17.83

Suponga que todas las interacciones son potencialmente importantes y con base en esto haga un análisis completo de los datos. Use valores *P* en la conclusión.

15.3 En un experimento metalúrgico se desea probar el efecto de cuatro factores y sus interacciones sobre la concentración (porcentaje por peso) de cierto compuesto particular de fósforo en el material de fundición. Las variables son *A*, porcentaje de fósforo en la refinación; *B*, porcentaje del material vuelto a fundir; *C*, tiempo de flujo; y *D*, tiempo de espera. Se varían los cuatro factores en un experimento factorial 2^4, con dos fundiciones tomadas de cada combinación de factores. Las 32 fundiciones se hicieron en orden aleatorio. Los datos se muestran en la siguiente tabla, y en la figura 15.8 de la página 610 se incluye la tabla del ANOVA. Analice los efectos de los factores y sus interacciones sobre la concentración del compuesto de fósforo.

Combinación de tratamientos	Peso % de compuesto de fósforo		
	Réplica 1	Réplica 2	Total
(1)	30.3	28.6	58.9
a	28.5	31.4	59.9
b	24.5	25.6	50.1
ab	25.9	27.2	53.1
c	24.8	23.4	48.2
ac	26.9	23.8	50.7
bc	24.8	27.8	52.6
abc	22.2	24.9	47.1
d	31.7	33.5	65.2
ad	24.6	26.2	50.8
bd	27.6	30.6	58.2
abd	26.3	27.8	54.1
cd	29.9	27.7	57.6
acd	26.8	24.2	51.0
bcd	26.4	24.9	51.3
abcd	26.9	29.3	56.2
Total	428.1	436.9	865.0

15.4 Se realizó un experimento preliminar para estudiar los efectos de cuatro factores y sus interacciones sobre la producción de la operación de cierta máquina. Se realizan dos corridas de cada una de las combinaciones de tratamientos para obtener una medida del error experimental puro. Se emplean dos niveles de cada factor y se obtienen los datos que se observan en la siguiente página. Pruebe todos los efectos principales y las interacciones a un nivel de significancia de 0.05. Saque sus conclusiones.

Fuente de variación	Efectos	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada	Valor P
Efecto principal:						
A	−1.2000	11.52	1	11.52	4.68	0.0459
B	−1.2250	12.01	1	12.01	4.88	0.0421
C	−2.2250	39.61	1	39.61	16.10	0.0010
D	1.4875	17.70	1	17.70	7.20	0.0163
Interacción de dos factores:						
AB	0.9875	7.80	1	7.80	3.17	0.0939
AC	0.6125	3.00	1	3.00	1.22	0.2857
AD	−1.3250	14.05	1	14.05	5.71	0.0295
BC	1.1875	11.28	1	11.28	4.59	0.0480
BD	0.6250	3.13	1	3.13	1.27	0.2763
CD	0.7000	3.92	1	3.92	1.59	0.2249
Interacción de tres factores:						
ABC	−0.5500	2.42	1	2.42	0.98	0.3360
ABD	1.7375	24.15	1	24.15	9.82	0.0064
ACD	1.4875	17.70	1	17.70	7.20	0.0163
BCD	−0.8625	5.95	1	5.95	2.42	0.1394
Interacción de cuatro factores:						
$ABCD$	0.7000	3.92	1	3.92	1.59	0.2249
Error		39.36	16	2.46		
Total		217.51	31			

Figura 15.8: Tabla ANOVA para el ejercicio 15.3.

Combinación de tratamientos	Réplica 1	Réplica 2
(1)	7.9	9.6
a	9.1	10.2
b	8.6	5.8
c	10.4	12.0
d	7.1	8.3
ab	11.1	12.3
ac	16.4	15.5
ad	7.1	8.7
bc	12.6	15.2
bd	4.7	5.8
cd	7.4	10.9
abc	21.9	21.9
abd	9.8	7.8
acd	13.8	11.2
bcd	10.2	11.1
abcd	12.8	14.3

15.5 En el estudio *An X-Ray Fluorescence Method for Analyzing Polybutadiene-Acrylic Acid (PBAA) Propellants* (Quarterly Reports, RK-TR-62-1, Army Ordnance Missile Command) se realizó un experimento para determinar si existe o no una diferencia significativa en la cantidad de aluminio obtenido en un análisis con ciertos niveles de ciertas variables de procesamiento. A continuación se presentan los datos.

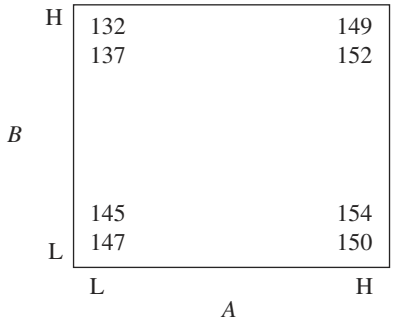
Observación	Estado físico	Tiempo de mezclado	Vel. de las aspas	Condición de nitrógeno	Aluminio
1	1	1	2	2	16.3
2	1	2	2	2	16.0
3	1	1	1	1	16.2
4	1	2	1	2	16.1
5	1	1	1	2	16.0
6	1	2	1	1	16.0
7	1	2	2	1	15.5
8	1	1	2	1	15.9
9	2	1	2	2	16.7
10	2	2	2	2	16.1
11	2	1	1	1	16.3
12	2	2	1	2	15.8
13	2	1	1	2	15.9
14	2	2	1	1	15.9
15	2	2	2	1	15.6
16	2	1	2	1	15.8

Las variables para los datos son:
A: tiempo de mezcla
nivel 1: 2 horas
nivel 2: 4 horas

B: Velocidad de las aspas
nivel 1: 36 rpm
nivel 2: 78 rpm
C: Condición de nitrógeno que pasa por el propulsor
nivel 1: seco
nivel 2: 72% de humedad relativa
D: estado físico del propulsor
nivel 1: no refinado
nivel 2: refinado

Analice los datos suponiendo que todas las interacciones de tres y cuatro factores son despreciables. Utilice un nivel de significancia de 0.05. Escriba un breve informe que resuma sus hallazgos.

15.6 Es importante estudiar el efecto de la concentración del reactivo y la tasa de alimentación de la viscosidad del producto de cierto proceso químico. La concentración del reactivo será el factor *A* a los niveles 15% y 25%. La tasa de alimentación será el factor *B* a niveles de 20 lb/h y 30 lb/h. El experimento implica 2 corridas experimentales en cada una de las cuatro combinaciones (L = bajo y H = alto). Las lecturas de la viscosidad son las siguientes.



a) Suponga un modelo que contiene dos efectos principales y una interacción y calcule los tres efectos. ¿Tiene usted alguna interpretación en este momento?
b) Realice un análisis de varianza y haga pruebas de interacción. Saque conclusiones.
c) Realice pruebas para los efectos principales y saque conclusiones finales acerca de la importancia de todos estos efectos.

15.7 Considere el ejercicio 15.3. Al investigador no sólo le interesa saber que las interacciones *AD*, *BC* y quizá *AB* son importantes, sino también su significado científico. Dibuje gráficas de interacción bidimensional para las tres e interprételas.

15.8 Considere nuevamente el ejercicio 15.3. Es frecuente que las interacciones de tres factores no sean significativas y, aun si lo fueran, serían difíciles de interpretar. La interacción *ABD* parece ser importante. Para hacer cierta interpretación dibuje dos gráficas de la interacción *AD*, una para $B = -1$ y otra para $B = +1$. A partir de la apariencia de éstas interprete la interacción *ABD*.

15.9 Considere el ejercicio 15.6. Utilice una escala de +1 y −1, para "alto" y "bajo", respectivamente, y calcule una regresión lineal múltiple con el modelo

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

con x_{1i}, = concentración del reactivo (−1, +1) y x_{2i} = tasa de alimentación (−1, +1).
a) Calcule los coeficientes de regresión.
b) ¿Cómo se relacionan los coeficientes b_1, b_2 y b_{12} con los efectos que encontró en el ejercicio 15.6*a*)?
c) En su análisis de regresión haga pruebas *t* sobre b_1, b_2 y b_{12}. ¿Cómo se relacionan estos resultados de la prueba con los del ejercicio 15.6*b*) y *c*)?

15.10 Considere el ejercicio 15.5. Calcule los 15 efectos y haga gráficas de probabilidad normal de los efectos.
a) ¿Parece válida la suposición de que las interacciones de tres y cuatro factores son despreciables?
b) ¿Los resultados de las gráficas del efecto son consistentes con lo que usted planteó sobre la importancia de los efectos principales y las interacciones de dos factores en su informe de resumen?

15.11 En Myers, Montgomery y Anderson-Cook (2009) se analiza un conjunto de datos para el que un ingeniero empleó un factorial 2^3 con el fin de estudiar los efectos de la velocidad de corte (*A*), la geometría de la herramienta (*B*) y el ángulo de corte (*C*) sobre la vida (en horas) de una máquina. Se eligen dos niveles de cada factor y se hacen pruebas dobles en cada punto del diseño en un orden aleatorio. A continuación se presentan los datos.

	A	***B***	***C***	Vida
(1)	−	−	−	22, 31
a	+	−	−	32, 43
b	−	+	−	35, 34
ab	+	+	−	35, 47
c	−	−	+	44, 45
ac	+	−	+	40, 37
bc	−	+	+	60, 50
abc	+	+	+	39, 41

a) Calcule los siete efectos. Con base en su magnitud, ¿cuál parece ser importante?
b) Haga un análisis de varianza y observe los valores *P*.
c) ¿Coinciden los resultados de los incisos *a* y *b*?

d) El ingeniero confía en que debe haber una interacción entre la velocidad y el ángulo de corte. Si esta interacción es significativa, dibuje una gráfica de la interacción y analice su significado desde el punto de vista de la ingeniería.

15.12 Considere el ejercicio 15.11 y suponga que hubo cierta dificultad experimental para hacer las corridas; que en realidad se tuvo que suspender todo el experimento después de sólo cuatro corridas. Como resultado, el experimento abreviado es dado por

	Vida
a	43
b	35
c	44
abc	39

Con sólo estas corridas los signos para los contrastes son dados por

	A	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+

Comente y determine si los contrastes son o no ortogonales. ¿Cuáles lo son y cuáles no? ¿Los efectos principales son ortogonales entre sí? En ese experimento abreviado (denominado *factorial fraccionario*) ¿es posible estudiar las interacciones de los efectos principales en forma independiente? ¿Se trataría de un experimento útil si estuviéramos convencidos de que las interacciones son despreciables? Explique su respuesta.

15.4 Experimentos factoriales en un ajuste de regresión

Hasta ahora hemos limitado el análisis de los datos para un factorial 2^k al método del análisis de varianza. La única referencia a un análisis alternativo se hizo en el ejercicio 15.9 de la página 611. De hecho, este ejercicio introduce gran parte del material que da origen a la presente sección. Hay situaciones en las que el ajuste de un modelo es importante **y en la que es posible controlar** los factores que se estudian. Por ejemplo, un biólogo podría querer estudiar el crecimiento de cierto tipo de alga en el agua, en cuyo caso sería muy útil un modelo que relacionara las unidades de algas como una función de la *cantidad de cierto contaminante*, y, digamos, del *tiempo*. Así, el estudio involucra un experimento factorial en un ambiente de laboratorio en el que los factores son la concentración del contaminante y el tiempo. Como se verá más adelante en esta sección, es posible ajustar un modelo más preciso si los factores están controlados en un arreglo factorial, para el que con frecuencia es útil elegir un factorial 2^k. En muchos procesos biológicos y químicos los niveles de las variables regresoras pueden y deberían controlarse.

Hay que recordar que el modelo de regresión empleado en el capítulo 12 se puede escribir con notación de matriz de la siguiente manera

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

La matriz $\mathbf{X}$ se denomina **matriz del modelo**. Suponga, por ejemplo, que se utiliza un experimento factorial 2^3 con las variables

Temperatura:	150°C	200°C
Humedad:	15%	20%
Presión (psi):	1000	1500

Los niveles familiares $+1$ y -1 se generan a través del siguiente centrado y escalado a *unidades de diseño*:

$$x_1 = \frac{\text{temperatura} - 175}{25}, \quad x_2 = \frac{\text{humedad} - 17.5}{2.5}, \quad x_3 = \frac{\text{presión} - 1250}{250}.$$

Como resultado, la matriz **X** se vuelve

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{c} \text{Identificación del diseño} \\ (1) \\ a \\ b \\ c \\ ab \\ ac \\ bc \\ abc \end{array}$$

Ahora se observa que los contrastes ilustrados y analizados en la sección 15.2 están relacionados directamente con los coeficientes de regresión. Observe que todas las columnas de la matriz **X** en el ejemplo 2^3 son *ortogonales*. Como resultado, el cálculo de los coeficientes de regresión que se describió en la sección 12.3 se convierte en

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\frac{1}{8}\mathbf{I}\right)\mathbf{X}'\mathbf{y}$$

$$= \frac{1}{8}\begin{bmatrix} a + ab + ac + abc + (1) + b + c + bc \\ a + ab + ac + abc - (1) - b - c - bc \\ b + ab + bc + abc - (1) - a - c - ac \\ c + ac + bc + abc - (1) - a - b - ab \end{bmatrix},$$

donde *a*, *ab*, etc., son medidas de la respuesta.

Ahora se observa que el concepto de *principales efectos calculados* que se enfatiza a lo largo de todo este capítulo con diseños factoriales 2^k, se relaciona con los coeficientes de un modelo de regresión ajustado cuando los factores son cuantitativos. De hecho, para un 2^k con, digamos, *n* corridas experimentales por punto del diseño, las relaciones entre los efectos y los coeficientes de regresión son como sigue:

$$\text{Efecto} = \frac{\text{contraste}}{2^{k-1}(n)}$$

$$\text{Coeficiente de regresión} = \frac{\text{contraste}}{2^k(n)} = \frac{\text{efecto}}{2}.$$

Esta relación debería tener sentido para el lector, ya que un coeficiente de regresión b_j es una tasa promedio del cambio en la respuesta *por cambio de unidad* en x_j. Por supuesto, cuando se va de -1 a $+1$ en x_j (de bajo a alto), la variable de diseño cambia en 2 unidades.

Ejemplo 15.2: Considere un experimento donde un ingeniero desea ajustar una regresión lineal del producto *y* contra el tiempo de retención x_1 y el tiempo de flexión x_2 en cierto sistema químico. Todos los demás factores se mantienen fijos. Los datos en las unidades naturales se incluyen en la tabla 15.8. Estime el modelo de regresión lineal múltiple.

***Solución*:** El modelo de regresión ajustado es

$$\hat{y} = b_0 + b_1x_1 + b_2x_2.$$

Tabla 15.8: Datos para el ejemplo 15.2

Tiempo de retención (hr)	Tiempo de flexión(hr)	Producto (%)
0.5	0.10	28
0.8	0.10	39
0.5	0.20	32
0.8	0.20	46

Las unidades de diseño son

$$x_1 = \frac{\text{tiempo de retención} - 0.65}{0.15}, \qquad x_2 = \frac{\text{tiempo de flexión} - 0.15}{0.05}$$

y la matriz $\mathbf{X}$ es

$$\begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} \end{array}$$

con los coeficientes de regresión

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \dfrac{(1) + a + b + ab}{4} \\ \dfrac{a + ab - (1) - b}{4} \\ \dfrac{b + ab - (1) - a}{4} \end{bmatrix} = \begin{matrix} 36.25 \\ 6.25 \\ 2.75 \end{matrix}.$$

Así, la ecuación de regresión de mínimos cuadrados es

$$\hat{y} = 36.25 + 6.25x_1 + 2.75x_2.$$

Este ejemplo ilustra el uso del experimento factorial de dos niveles en un ajuste de regresión. Las cuatro corridas experimentales en el diseño 2^2 se usaron para obtener una ecuación de regresión, con la interpretación evidente de los coeficientes de regresión. El valor $b_1 = 6.25$ representa el incremento estimado en la respuesta (porcentaje de producción) por cambio en la *unidad de diseño* (0.15 horas) en el tiempo de retención. El valor $b_2 = 2.75$ representa una tasa de cambio similar para el tiempo de flexión. ■

Interacción en el modelo de regresión

Los contrastes de interacción que se estudiaron en la sección 15.2 tienen interpretaciones definidas en el contexto de la regresión. De hecho, las interacciones se explican en los modelos de regresión en términos de producto. Esto se ilustra en el ejemplo 15.2, en donde el modelo con interacción es

$$y = b_0 + b_1x_1 + b_2x_2\, b_{12}x_1x_2$$

con b_0, b_1 y b_2, como antes, y

$$b_{12} = \frac{ab + (1) - a - b}{4} = \frac{46 + 28 - 39 - 32}{4} = 0.75.$$

Así, la ecuación de regresión que expresa dos *efectos principales lineales* e interacción, es

$$\hat{y} = 36.25 + 6.25x_1 + 2.75x_2 + 0.75x_1x_2.$$

El contexto de la regresión proporciona un marco de referencia mediante el cual el lector debería entender mejor la ventaja de la ortogonalidad de que goza el factorial 2^k. En la sección 15.2 se analizaron las ventajas de la ortogonalidad desde el punto de vista del *análisis de varianza* de los datos en un experimento factorial 2^k. Se señaló que la ortogonalidad entre los efectos conduce a la independencia entre las sumas de cuadrados. Desde luego, la presencia de variables de regresión no descarta el uso del análisis de varianza. De hecho, las pruebas *f* se llevan a cabo tal como se describió en la sección 15.2. No obstante, se debe hacer una distinción. En el caso del ANOVA las hipótesis surgen de medias poblacionales, mientras que en el caso de la regresión las hipótesis implican coeficientes de regresión.

Por ejemplo, considere el diseño experimental del ejercicio 15.2 de la página 609. Cada factor es continuo. Suponga que los niveles son

A (x_1):	20%	50%
B (x_2):	5 lb/sec	10 lb/sec
C (x_3):	5	5.5

y que se tiene, para los niveles de diseño,

$$x_1 = \frac{\%\ \text{sólidos} - 30}{10}, \qquad x_2 = \frac{\text{tasa de flujo} - 7.5}{2.5}, \qquad x_3 = \frac{\text{pH} - 5.25}{0.25}.$$

Suponga que es de interés ajustar un modelo de regresión múltiple, en el cual se considerarán todos los coeficientes lineales y las interacciones disponibles. Además, el ingeniero desea obtener información acerca de cuáles niveles del factor *maximizarán* la purificación, es decir, maximizar la respuesta. Este problema es el tema del estudio de caso 15.2.

Estudio de caso 15.2: **Experimento de purificación del carbón:**[1] La figura 15.9 representa una salida de resultados comentados del análisis de regresión del modelo ajustado

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_2 + b_{23}x_2x_{23} + b_{123}x_1x_2x_3,$$

donde x_1, x_2 y x_3 representan al porcentaje de sólidos, la tasa de flujo y el pH del sistema, respectivamente. Se utilizó la función PROC REG del sistema de cómputo *SAS*.

Observe los estimados del parámetro, el error estándar y los valores *P* en la salida de resultados por computadora. Los estimados del parámetro representan los coeficientes del modelo. Todos ellos son significativos, excepto el término x_2x_3 (interacción *BC*). También observe que los residuales, los intervalos de confianza y los intervalos de predicción aparecen como se presentaron en el material sobre regresión de los capítulos 11 y 12.

El lector puede usar los valores de los coeficientes del modelo y los valores pronosticados en la salida de resultados por computadora para asegurarse de que la combinación de los factores dé como resultado la **mayor eficiencia de pureza**. El factor *A* (porcentaje de sólidos circulados) tiene un coeficiente positivo alto, lo cual sugiere un valor elevado para el porcentaje de sólidos. Además, se sugiere un valor bajo para el factor *C* (pH del tanque). Aunque el coeficiente del efecto principal *B* (tasa de flujo

[1]Véase el ejercicio 15.2.

```
Dependent Variable: Y
                    Analysis of Variance
                         Sum of      Mean
Source            DF    Squares    Square  F Value  Pr > F
Model              7  490.23499  70.03357   254.43  <.0001
Error              8    2.20205   0.27526
Corrected Total   15  492.43704
Root MSE         0.52465    R-Square     0.9955
Dependent Mean  12.75188    Adj R-Sq     0.9916
Coeff Var        4.11429
                  Parameter Estimates
                Parameter  Standard
Variable   DF   Estimate     Error    t Value  Pr > |t|
Intercept   1   12.75188   0.13116      97.22    <.0001
A           1    4.71938   0.13116      35.98    <.0001
B           1    0.86563   0.13116       6.60    0.0002
C           1   -1.41563   0.13116     -10.79    <.0001
AB          1   -0.59938   0.13116      -4.57    0.0018
AC          1   -0.52813   0.13116      -4.03    0.0038
BC          1    0.00562   0.13116       0.04    0.9668
ABC         1    2.23063   0.13116      17.01    <.0001
    Dependent Predicted    Std Error
Obs  Variable     Value Mean  Predict    9 5% CL Mean       95% CL Predict Residual
  1    4.6500    5.2300        0.3710   4.3745   6.0855   3.7483   6.7117  -0.5800
  2   21.4200   21.3850        0.3710  20.5295  22.2405  19.9033  22.8667   0.0350
  3   12.6600   12.6100        0.3710  11.7545  13.4655  11.1283  14.0917   0.0500
  4   18.2700   17.4450        0.3710  16.5895  18.3005  15.9633  18.9267   0.8250
  5    7.9300    7.9050        0.3710   7.0495   8.7605   6.4233   9.3867   0.0250
  6   13.1800   13.0250        0.3710  12.1695  13.8805  11.5433  14.5067   0.1550
  7    6.5100    6.3850        0.3710   5.5295   7.2405   4.9033   7.8667   0.1250
  8   18.2300   18.0300        0.3710  17.1745  18.8855  16.5483  19.5117   0.2000
  9    5.8100    5.2300        0.3710   4.3745   6.0855   3.7483   6.7117   0.5800
 10   21.3500   21.3850        0.3710  20.5295  22.2405  19.9033  22.8667  -0.0350
 11   12.5600   12.6100        0.3710  11.7545  13.4655  11.1283  14.0917  -0.0500
 12   16.6200   17.4450        0.3710  16.5895  18.3005  15.9633  18.9267  -0.8250
 13    7.8800    7.9050        0.3710   7.0495   8.7605   6.4233   9.3867  -0.0250
 14   12.8700   13.0250        0.3710  12.1695  13.8805  11.5433  14.5067  -0.1550
 15    6.2600    6.3850        0.3710   5.5295   7.2405   4.9033   7.8667  -0.1250
 16   17.8300   18.0300        0.3710  17.1745  18.8855  16.5483  19.5117  -0.2000
```

Figura 15.9: Lista de resultados del *SAS* para los datos del estudio de caso 15.2.

del polímero) es positivo, el coeficiente positivo elevado de $x_1x_2x_3$ (ABC) sugiere que la tasa de flujo debería estar en el nivel bajo para aumentar la eficiencia. De hecho, el modelo de regresión generado en la salida de resultados por computadora del *SAS* sugiere que la combinación de factores que podrían producir resultados óptimos, o quizá sugerir experimentos futuros, es dada por

A: nivel alto
B: nivel bajo
C: nivel bajo

15.5 El diseño ortogonal

En situaciones experimentales en las que es apropiado ajustar modelos que son lineales en las variables de diseño y que posiblemente impliquen interacciones o términos de producto, el *diseño ortogonal* de dos niveles, o arreglo ortogonal, plantea algunas ventajas. Por diseño ortogonal nos referimos a uno en el que hay ortogonalidad entre las columnas de la matriz **X**. Considere la matriz **X** para el factorial 2^2 del ejemplo 15.2. Observe que las tres columnas son mutuamente ortogonales. La matriz **X** del factorial 2^3 también contiene columnas ortogonales. El factorial 2^3 con interacciones produciría una matriz **X** del tipo

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \quad & x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 & x_1x_2x_3 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

La descripción de los grados de libertad es

Fuente	**g.l.**	
Regresión	3	
Falta de ajuste	4	$(x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$
Error (puro)	8	
Total	15	

Los ocho grados de libertad para el error puro se obtienen a partir de las *corridas duplicadas* en cada punto del diseño. Los grados de libertad de la falta de ajuste podrían considerarse como la diferencia entre el número de puntos de diseño distintos y el número total de términos en el modelo; en este caso hay ocho puntos y cuatro términos en el modelo.

Error estándar de los coeficientes y pruebas *T*

En las secciones anteriores vimos cómo el diseñador de un experimento puede aprovechar el concepto de ortogonalidad para diseñar un experimento de regresión con coeficientes que obtienen una varianza mínima sobre la base del costo. Debemos ser capaces de utilizar el material sobre la regresión que se expuso en la sección 12.4 para calcular estimados de las varianzas de los coeficientes y, con ello, los errores estándar. También resulta de interés observar la relación entre el estadístico *t* de un coeficiente y el estadístico *F* descrito e ilustrado en capítulos anteriores.

En la sección 12.4 vimos que las varianzas y las covarianzas de los coeficientes aparecen en A^{-1}, o, en términos de la notación actual, la *matriz de varianza-covarianza* de coeficientes es

$$\sigma^2 A^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

En el caso del experimento factorial 2^k las columnas de **X** son mutuamente ortogonales,

lo que impone una estructura muy especial. En general, para 2^k se puede escribir

$$\begin{array}{cccccccc} & x_1 & x_2 & \cdots & x_k & x_1x_2 & \cdots \\ \mathbf{X} = [\mathbf{1} & \pm\mathbf{1} & \pm\mathbf{1} & \cdots & \pm\mathbf{1} & \pm\mathbf{1} & \cdots], \end{array}$$

donde cada columna contiene 2^k o 2^kn entradas, donde n es el número de réplicas de las corridas en cada punto del diseño. Así, la formación de $\mathbf{X'X}$ lleva a

$$\mathbf{X'X} = 2^k n\mathbf{I}_p,$$

donde $\mathbf{I}$ es la matriz de identidad de la dimensión p, el número de parámetros del modelo.

Ejemplo 15.3: Considere un diseño factorial 2^3 con corridas por duplicado que se ajusta al modelo

$$E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3.$$

Proporcione expresiones para los errores estándar de los estimados de mínimos cuadrados de b_0, b_1, b_2, b_3, b_{12}, b_{13} y b_{23}.

***Solución*:**

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

considerando cada unidad como *repetida,* es decir, considerando que cada observación está duplicada. Como resultado,

$$\mathbf{X'X} = 16\mathbf{I}_7.$$

Por consiguiente,

$$(\mathbf{X'X})^{-1} = \frac{1}{16}\mathbf{I}_7.$$

A partir de lo anterior debe quedar claro que las varianzas de todos los coeficientes para un factorial 2^k con n corridas en cada punto de diseño son

$$\text{Var}(b_j) = \frac{\sigma^2}{2^k n},$$

y, desde luego, todas las covarianzas son iguales a cero. Como resultado, los errores estándar de los coeficientes se calculan como

$$s_{b_j} = s\sqrt{\frac{1}{2^k n}},$$

donde s se calcula por medio de la raíz cuadrada del cuadrado medio del error que se espera obtener a partir de una réplica adecuada. Así, en nuestro caso con 2^3,

$$s_{b_j} = s\left(\frac{1}{4}\right).$$

Ejemplo 15.4: Considere el experimento metalúrgico del ejercicio 15.3 de la página 609. Suponga que el modelo ajustado es

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4.$$

¿Cuáles son los errores estándar de los coeficientes de regresión de los mínimos cuadrados?

***Solución*:** Los errores estándar de todos los coeficientes para el factorial 2^k son iguales, y son

$$s_{b_j} = s\sqrt{\frac{1}{2^k n}},$$

que en este ejemplo es

$$s_{b_j} = s\sqrt{\frac{1}{(16)(2)}}.$$

En este caso el cuadrado medio del error puro es dado por $s^2 = 2.46$ (16 grados de libertad). Entonces,

$$s_{b_j} = 0.28.$$

Los errores estándar de los coeficientes se usan para construir estadísticos t de todos los coeficientes. Estos valores t se relacionan con los estadísticos F del análisis de varianza. Ya se demostró que un estadístico F sobre un coeficiente, usando el factorial 2^k, es

$$f = \frac{(\text{contraste})^2}{(2^k n)s^2}.$$

Ésta es la forma del estadístico F de la página 610 para el experimento metalúrgico (ejercicio 15.3). Es fácil comprobar que si se escribe

$$t = \frac{b_j}{s_{b_j}}, \qquad \text{donde} \qquad b_j = \frac{\text{contraste}}{2^k n},$$

entonces

$$t^2 = \frac{(\text{contraste})^2}{s^2 2^k n} = f.$$

Como resultado, se mantiene la relación acostumbrada entre los estadísticos t sobre los coeficientes y los valores F. Como era de esperarse, la única diferencia entre utilizar t y F para evaluar la significancia radica en el hecho de que el estadístico t indica el signo o la dirección del efecto del coeficiente.

Parecería que el plan del factorial 2^k se adapta a muchas situaciones prácticas a las cuales se ajustan modelos de regresión. Puede incluir términos lineales y de interacción, lo que proporciona estimados óptimos de todos los coeficientes (desde un punto de vista de la varianza). Sin embargo, cuando k es grande, el número de puntos del diseño requerido es muy grande. A menudo es posible utilizar partes del diseño total y aun así conservar la ortogonalidad, con todas sus ventajas. En la sección 15.6 se estudian esos diseños.

Una mirada más cercana a la propiedad de ortogonalidad del factorial 2^k

Ya vimos que para el caso del factorial 2^k toda la información que obtiene el analista sobre los efectos y las interacciones principales aparece en forma de contrastes. Estas "$2^k - 1$ piezas de información" conllevan un solo grado de libertad cada una y son independientes entre sí. En un análisis de varianza se manifiestan como *efectos*; mientras que si se construye un modelo de regresión, los efectos que resultan son coeficientes de regresión, aparte de un factor de 2. Con cada forma de análisis es posible hacer pruebas de significancia y la prueba *t* para un efecto dado es la misma en términos numéricos que para el coeficiente de regresión correspondiente. En el caso del ANOVA son importantes la selección de las variables y la interpretación científica de las interacciones; en tanto que en el caso de un análisis de regresión se usa un modelo para predecir la respuesta y/o determinar cuáles combinaciones de factores o niveles son las óptimas, por ejemplo, maximizar la producción o la eficiencia de la purificación, como en el estudio de caso 15.2.

Resulta que la propiedad de ortogonalidad es importante, ya sea que se trate de un ANOVA o de una regresión. La ortogonalidad entre las columnas de X, la matriz del modelo en, digamos, el ejemplo 15.3, ofrece condiciones especiales que tienen un impacto importante sobre los **efectos de la varianza** o los **coeficientes de regresión**. De hecho, ya es evidente que el diseño ortogonal da como resultado la igualdad de varianza para todos los efectos o coeficientes. Es así como, para propósitos de estimación o de prueba, la precisión es la misma para todos los coeficientes, los efectos principales o las interacciones. Además, si el modelo de regresión sólo contiene términos lineales, por lo cual sólo los efectos principales son de interés, las condiciones siguientes dan como resultado la minimización de las varianzas de todos los efectos, o, en forma correspondiente, de los coeficientes de regresión de primer orden.

Condiciones para varianzas mínimas de los coeficientes

Si el modelo de regresión contiene términos no mayores de primer orden, y si los rangos de las variables son dados por $x_j \in [-1, +1]$ para $j = 1, 2, \ldots, k$, entonces $\text{Var}(b_j)/\sigma^2$, para $j = 1, 2, \ldots, k$, se minimiza si el diseño es ortogonal y todos los niveles x_i del diseño son ± 1 para $i = 1, 2, \ldots, k$.

Así, en términos de los coeficientes del modelo o los efectos principales, la ortogonalidad en el 2^k es una propiedad muy deseable.

Otro método para lograr una mejor comprensión del "balance" proporcionado por el factorial 2^3 consiste en observar la situación mediante una gráfica. En la figura 15.10 se aprecia cada uno de los contrastes ortogonales y, por lo tanto, mutuamente independientes. En las gráficas se comparan los planos de los cuadrados cuyos vértices contienen las respuestas etiquetadas con "+" con las que tienen el signo "−". Las que aparecen en el inciso *a* presentan contrastes para efectos principales y deberían ser evidentes para el lector. Las del inciso *b* presentan los planos determinados por los vértices "+" y "−" para los tres contrastes de interacción de dos factores. En el inciso *c* se aprecia la representación geométrica de los contrastes para la interacción de tres factores (*ABC*).

Corridas centrales con diseños factoriales 2^k

En la situación en que se aplica el diseño 2^k con variables **continuas** de diseño y se busca ajustar un modelo de regresión lineal, el uso de réplicas de corridas en el **diseño central** puede ser sumamente útil. De hecho, además de las ventajas que se analizarán a continuación, la mayoría de los científicos e ingenieros considerarían que las corridas

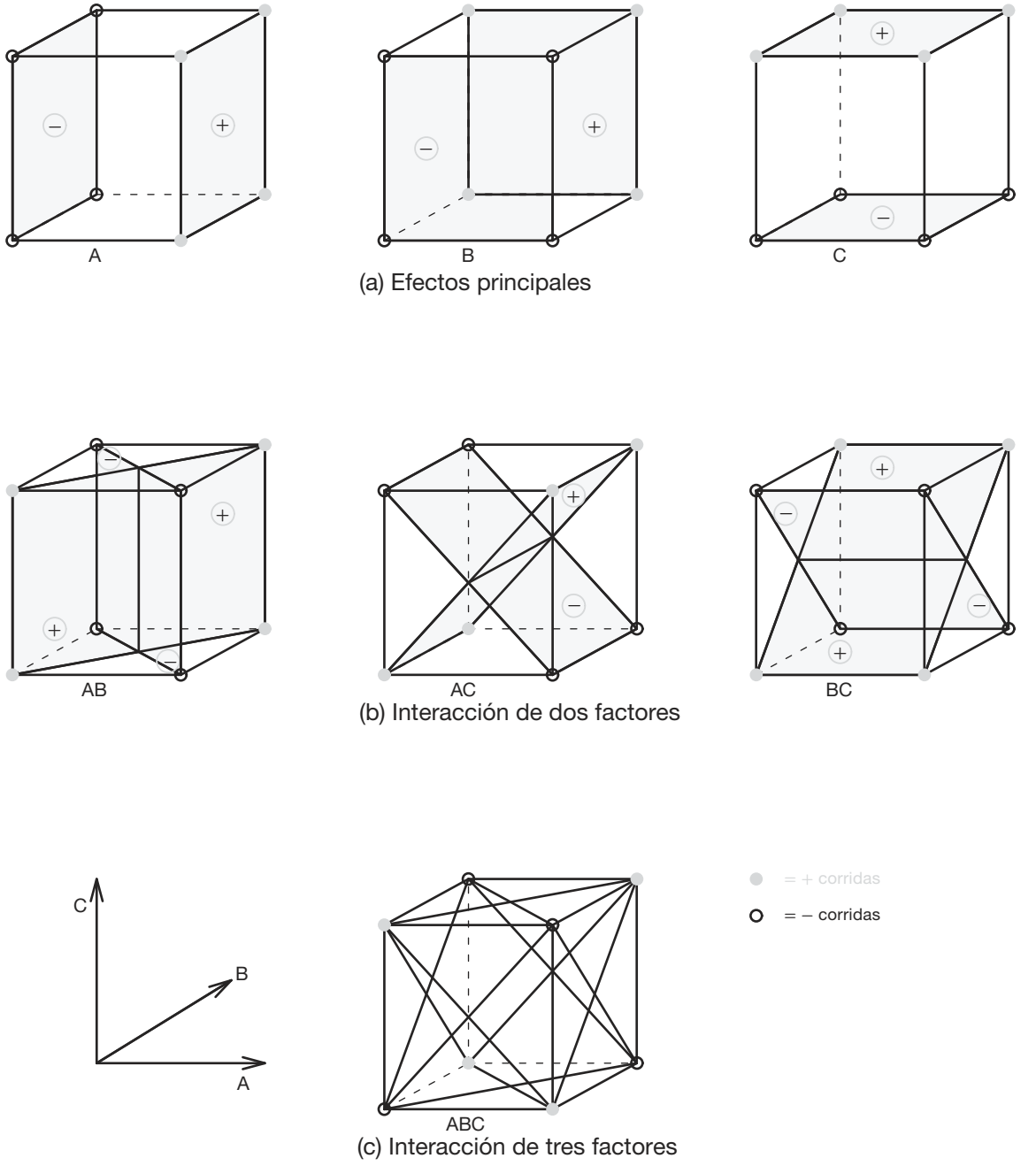


Figura 15.10: Presentación geométrica de los contrastes para el diseño factorial 2^3.

centrales, es decir, las corridas en $x_i = 0$ para $i = 1, 2, ..., k$, no sólo son una práctica razonable sino que además son interesantes. En muchas áreas de aplicación del diseño 2^k el científico desea determinar si sería benéfico pasar a otra región de interés en los factores. En muchos casos el centro, es decir, el punto (0, 0,..., 0) en los factores codificados, con frecuencia representa las condiciones de operación actuales del proceso, o al menos aquellas condiciones que se consideran "óptimas para el momento". Por lo tanto, a menudo el científico requerirá datos sobre la respuesta central.

Corridas centrales y falta de ajuste

Además del atractivo del aumento del diseño 2^k con corridas centrales, otra de sus ventajas consiste en que se relaciona con la clase de modelo que se ajusta a los datos. Considere, por ejemplo, el caso con $k = 2$ que se ilustra en la figura 15.11.

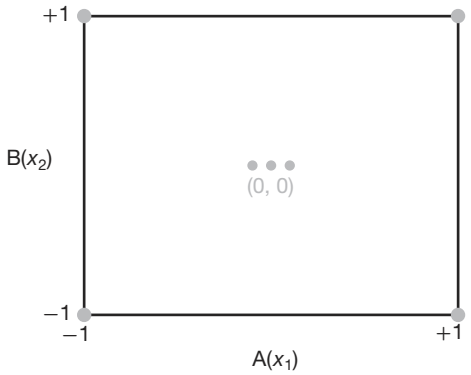


Figura 15.11: Un diseño 2^2 con corridas centrales.

Queda claro que *sin las corridas centrales* los términos del modelo son la intersección, x_1, x_2, x_1x_2. Esto explica los cuatro grados de libertad del modelo producidos por los cuatro puntos del diseño, además de cualquier réplica. Como cada factor tiene información de respuesta disponible *sólo en dos ubicaciones* $\{-1, +1\}$, no es posible incluir términos "puros" de curvatura de segundo orden en el modelo, es decir, x_1^2 o x_2^2. Sin embargo, la información en (0, 0) produce un grado de libertad adicional del modelo. Si bien este importante grado de libertad no permite que ni x_1^2 ni x_2^2 se empleen en el modelo, sí permite probar la significancia de una combinación lineal de x_1^2 y x_2^2. Entonces, para n_c corridas centrales, hay $n_c - 1$ grados de libertad disponibles para réplicas o para el error "puro". Esto permite un estimado de σ^2 para probar los términos del modelo y la significancia del único grado de libertad para la **falta de ajuste cuadrático**. El concepto aquí es muy similar al que se describió en el material sobre la falta de ajuste del capítulo 11.

Para entender por completo cómo funciona la prueba de falta de ajuste suponga que para $k = 2$ el **modelo verdadero** contiene todo el complemento de segundo orden de los términos, incluyendo a x_1^2 y x_2^2. En otras palabras,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2.$$

Ahora, considere el contraste

$$\bar{y}_f = \bar{y}_0$$

donde $\bar{y}_f$ es la respuesta promedio de las ubicaciones factoriales y $\bar{y}_0$ es la respuesta promedio en el punto central. Es fácil demostrar (véase el ejercicio de repaso 15.46) que

$$E(\bar{y}_f - \bar{y}_0) = \beta_{11} + \beta_{22},$$

y, en efecto, para el caso general con k factores,

$$E(\bar{y}_f - \bar{y}_0) = \sum_{i=1}^{k} \beta_{ii}.$$

Como resultado, la prueba de falta de ajuste es una prueba t simple (o $F = t^2$) con

$$t_{n_c-1} = \frac{\bar{y}_f - \bar{y}_0}{s_{\bar{y}_f - \bar{y}_0}} = \frac{\bar{y}_f - \bar{y}_0}{\sqrt{CME\,(1/n_f + 1/n_c)}},$$

donde n_f es el número de puntos factoriales y *CME* sólo es la varianza muestral de los valores de la respuesta en (0, 0, . . . , 0).

Ejemplo 15.5: Este ejemplo se tomó de Myers, Montgomery y Anderson-Cook (2009). Un ingeniero químico trata de modelar la conversión porcentual en un proceso. Hay dos variables de interés, el tiempo de reacción y la temperatura de reacción. En un intento por llegar al modelo apropiado se realiza un experimento preliminar en un factorial 2^2 usando la región actual de interés en el tiempo y temperatura de reacción. Se hicieron corridas únicas en cada uno de los cuatro puntos factoriales, y cinco corridas en el centro del diseño con el fin de poder realizar una prueba de falta de ajuste para la curvatura. En la figura 15.12 se presenta la región del diseño y las corridas experimentales sobre el producto.

Las lecturas del tiempo y la temperatura en el centro son, desde luego, 35 minutos y 145°C. Los estimados de los efectos principales y el coeficiente de interacción único se calculan mediante contrastes, igual que antes. Las corridas en el centro **no intervienen en el cálculo de b_1, b_2 y b_{12}**. Esto debería ser razonable para el lector. La intersección es sólo $\bar{y}$ para todo el experimento. Este valor es $\bar{y} = 40.4444$. Los errores estándar se calculan usando los elementos de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$, como ya se expuso. Para este caso,

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} 1 & x_1 & x_2 & x_1x_2 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

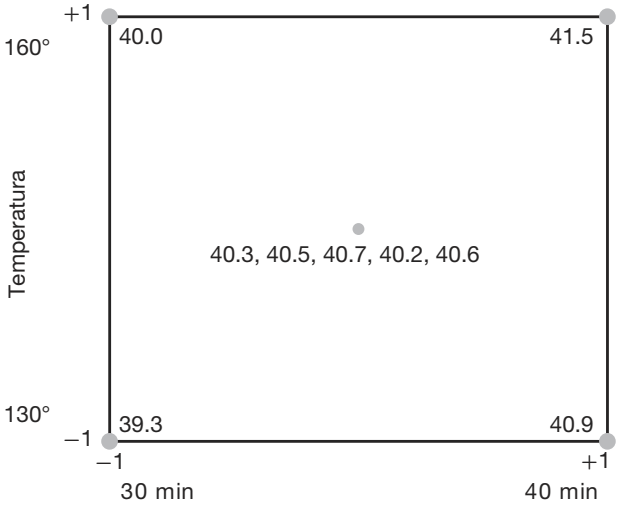


Figura 15.12: Factorial 2^2 con 5 corridas en el centro.

Después de hacer los cálculos se obtiene

$$
\begin{array}{llll}
b_0 = 40.4444, & b_1 = 0.7750, & b_2 = 0.3250, & b_{12} = -0.0250, \\
s_{b_0} = 0.06231, & s_{b_1} = 0.09347, & s_{b_2} = 0.09347, & s_{b_{12}} = 0.09347, \\
t_{b_0} = 649.07, & t_{b_1} = 8.29, & t_{b_2} = 3.48, & t_{b_{12}} = -0.27 \quad (P = 0.800).
\end{array}
$$

El contraste $\bar{y}_f - \bar{y}_0 = 40.425 - 40.46 = -0.035$ y el estadístico t que *prueba la curvatura* son dados por

$$t = \frac{40.425 - 40.46}{\sqrt{0.0430(1/4 + 1/5)}} = 0.251 \qquad (P = 0.814).$$

Como resultado, parece que el modelo apropiado debería contener sólo términos de primer orden (además de la intersección). ▟

Una mirada intuitiva a la prueba de curvatura

Si se considera el caso sencillo con una sola variable de diseño con corridas en -1 y $+1$ debe quedar claro que la respuesta promedio en -1 y $+1$ debe estar cerca de la respuesta en 0, el centro, si el modelo es de primer orden. Cualquier desviación sugeriría, con seguridad, curvatura. Esto se puede extender fácilmente a dos variables. Considere la figura 15.13.

La figura muestra el plano sobre y que pasa a través de los valores de y de los puntos factoriales. Éste es el plano que representaría el ajuste perfecto para el modelo que contiene x_1, x_2 y x_1x_2. Si el modelo no contiene curvatura cuadrática, es decir, $\beta_{11} = \beta_{22} = 0$, se esperaría que la respuesta en (0, 0) esté en el plano o cerca del mismo. Si la respuesta estuviera lejos del plano, como ocurre en la figura 15.13, entonces se podría ver en forma gráfica que la curvatura cuadrática está presente.

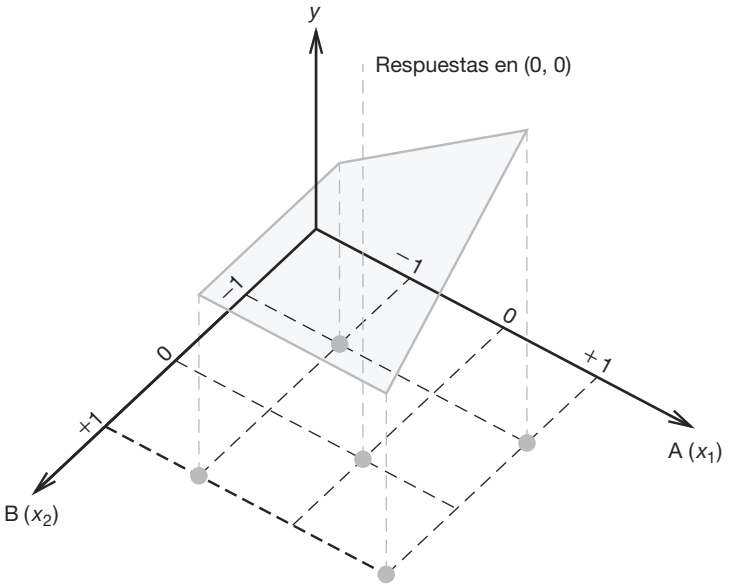


Figura 15.13: Factorial 2^2 con corridas en (0, 0).

Ejercicios

15.13 Considere un experimento 2^5 donde se realizan corridas experimentales sobre 4 máquinas diferentes. Use las máquinas como bloques y suponga que todos los efectos principales y las interacciones de dos factores son importantes.

a) ¿Cuáles corridas se harían sobre cada una de las 4 máquinas?

b) ¿Cuáles efectos se confunden con los bloques?

15.14 En un experimento descrito en Myers, Montgomery y Anderson-Cook (2009) se buscan las condiciones óptimas para almacenar semen de bovinos con el fin de obtener la supervivencia máxima. Las variables son el porcentaje de citrato de sodio, el porcentaje de glicerol y el tiempo de equilibrio en horas. La respuesta es el porcentaje de supervivencia de los espermatozoides móviles. Los niveles naturales se encuentran en la referencia mencionada. A continuación se presentan los datos con los niveles codificados para la parte factorial del diseño y las corridas centrales.

a) Ajuste un modelo de regresión lineal con los datos y determine cuáles términos lineales y de interacción son significativos. Suponga que la interacción $x_1x_2x_3$ es despreciable.

b) Pruebe la falta de ajuste cuadrático y comente la respuesta.

x_1, % de citrato de sodio	x_2, % de glicerol	x_3 Tiempo de equilibrio	% Supervivencia
−1	−1	−1	57
1	−1	−1	40
−1	1	1	19
1	1	1	40
−1	−1	−1	54
1	−1	−1	41
−1	1	1	21
1	1	1	43
0	0	0	63
0	0	0	61

15.15 Los productores de petróleo están interesados en aleaciones de níquel que sean fuertes y resistentes a la corrosión. Se realizó un experimento en el que se comparó del límite elástico especímenes elásticos de aleaciones de níquel cargados en una solución de ácido sulfúrico saturada con disulfuro de carbón. Se compararon dos aleaciones; una con 75% de níquel y otra con 30% de níquel. Se probaron las aleaciones en dos tiempos de carga diferentes, de 25 y 50 días. Se realizó un factorial 2^3 con los factores siguientes:

% de ácido sulfúrico: 4%, 6% (x_1)

tiempo de carga: 25 días, 50 días (x_2)

composición del níquel: 30%, 75% (x_3)

Se preparó un espécimen para cada una de las 8 condiciones. Como los ingenieros no estaban seguros de la naturaleza del modelo, es decir, de si se necesitarían o no términos cuadráticos, incorporaron un tercer nivel (intermedio) y realizaron 4 corridas centrales utilizando 4 especímenes con ácido sulfúrico al 5%, 37.5 días y una composición de níquel de 52.5%. A continuación se incluyen las resistencias en kilogramos por pulgada cuadrada.

	Tiempo de carga			
	25 días		**50 días**	
Composición	**Ácido sulfúrico**		**Ácido sulfúrico**	
del níquel	**4%**	**6%**	**4%**	**6%**
75%	52.5	56.5	47.9	47.2
30%	50.2	50.8	47.4	41.7

Las corridas centrales produjeron las siguientes resistencias:

51.6, 51.4, 52.4, 52.9

a) Haga pruebas para determinar cuáles efectos principales e interacciones deberían incluirse en el modelo ajustado.

b) Pruebe para la curvatura cuadrática.

c) Si la curvatura cuadrática es significativa, ¿cuántos puntos de diseño adicionales se necesitan para determinar cuáles términos cuadráticos deberían incluirse en el modelo?

15.16 Suponga que es posible llevar a cabo una réplica del experimento del ejercicio 15.13.

a) ¿Una segunda réplica del esquema de bloques del ejercicio 15.13 sería la mejor opción?

b) Si la respuesta del inciso *a* es negativa, proporcione el diseño de una mejor opción para la segunda réplica.

c) ¿Qué concepto utilizó en la elección del diseño?

15.17 Considere la figura 15.14, que representa un factorial 2^2 con 3 corridas centrales. Si la curvatura cuadrática es significativa, ¿cuáles otros puntos de diseño seleccionaría, que permitieran estimar los términos x_1^2 y x_2^2? Explique su respuesta.

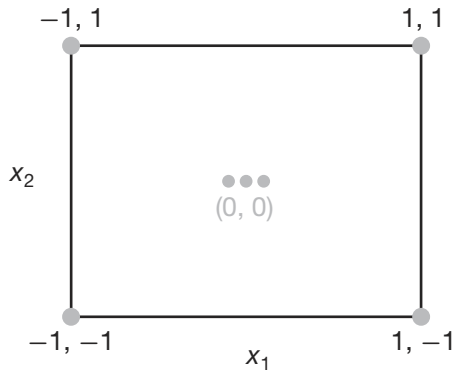


Figura 15.14: Gráfica para el ejercicio 15.17.

15.6 Experimentos factoriales fraccionarios

El experimento factorial 2^k se puede volver muy demandante, en términos del número de unidades experimentales requeridas, cuando el valor de k es grande. Una de las ventajas reales de este plan experimental es que permite un grado de libertad para cada interacción. Sin embargo, en muchas situaciones experimentales se sabe que ciertas interacciones son despreciables, por lo que sería un desperdicio de esfuerzo experimental utilizar el experimento factorial completo. De hecho, el experimentador podría tener limitaciones económicas que le impidan hacer observaciones de todas las combinaciones 2^k de tratamientos. Cuando k es grande, a menudo se puede usar un **experimento factorial**

fraccionario donde quizás sea posible llevar a cabo la mitad, un cuarto o incluso un octavo del plan factorial total.

Construcción de la fracción de $\frac{1}{2}$

La construcción del diseño de media réplica es idéntica a la asignación del experimento factorial 2^k en dos bloques. Se comienza por seleccionar un contraste de definición que se sacrificará por completo. Luego se construyen los dos bloques en concordancia y se elige cualquiera de ellos como plan experimental.

A menudo la fracción de $\frac{1}{2}$ de un factorial 2^k se conoce como diseño 2^{k-1}, el cual indica el número de puntos de diseño. El primer ejemplo de un diseño 2^{k-1} será uno de $\frac{1}{2}$ o uno de 2^3 o uno de 2^{3-1}. En otras palabras, el científico o el ingeniero no puede usar el complemento completo, es decir, todo el diseño 2^3 con 8 puntos de diseño, por lo que debe apelar a un diseño con sólo cuatro puntos de diseño. La pregunta es la siguiente: de los puntos de diseño (1), *a*, *b*, *ab*, *ac*, *c*, *bc* y *abc*, ¿cuáles son los cuatro puntos de diseño que producirán el diseño más útil? La respuesta, junto con los conceptos importantes relacionados, aparece en la tabla de signos + y − que muestra los contrastes para el diseño 2^3 completo. Considere la tabla 15.9.

Tabla 15.9: Contrastes para los siete efectos disponibles en el caso de un experimento factorial 2^3

	Combinación de tratamientos	Efectos							
		I	***A***	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
2^{3-1}	*a*	+	+	−	−	−	−	+	+
	b	+	−	+	−	−	+	−	+
	c	+	−	−	+	+	−	−	+
	abc	+	+	+	+	+	+	+	+
2^{3-1}	*ab*	+	+	+	−	+	−	−	−
	ac	+	+	−	+	−	+	−	−
	bc	+	−	+	+	−	−	+	−
	(1)	+	−	−	−	+	+	+	−

Observe que las dos fracciones $\frac{1}{2}$ son {*a, b, c, abc*} y {*ab, ac*, *bc,* (1)}. Observe también en la tabla 15.9 que en ambos diseños *ABC* no tiene contraste, pero todos los demás efectos sí lo tienen. En una de las fracciones se tiene que *ABC* contiene todos los signos + y en la otra fracción el efecto *ABC* contiene todos los signos −. Como resultado, se dice que el diseño de la parte superior de la tabla es descrito por $\boldsymbol{ABC = I}$, y el de la parte inferior por $\boldsymbol{ABC = -I}$. La interacción ***ABC*** se denomina **generador del diseño**, y $ABC = I$ (o $ABC = -I$ para el segundo diseño) recibe el nombre de **relación definitoria**.

Alias en el 2^{3-1}

Si nos centramos en el diseño $ABC = I$ (el 2^{3-1} superior), es evidente que seis efectos contienen contrastes. Esto produce la apariencia inicial de que todos los *efectos* se pueden estudiar por separado de *ABC*. Sin embargo, el lector recordará que con sólo cuatro puntos de diseño, incluso si se replican, los grados de libertad disponibles (además del error experimental) son

Términos del modelo de regresión	3
Intersección	1
	4

Un análisis más detallado sugiere que los siete efectos no son ortogonales y que cada contraste está representado en otro efecto. De hecho, si se emplea el símbolo $\equiv$ para denotar **contrastes idénticos**, se tiene que

$$A \equiv BC; \qquad B \equiv AC; \qquad C \equiv AB.$$

Como resultado, dentro de un par no es posible estimar un efecto independiente de su "socio" alias. Los efectos

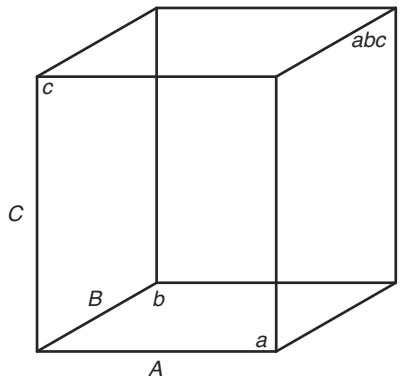
$$A = \frac{a + abc - b - c}{2} \qquad \text{y} \qquad BC = \frac{a + abc - b - c}{2}$$

producirán el mismo resultado numérico, de manera que contienen la misma información. De hecho, con frecuencia se dice que **comparten un grado de libertad**. En realidad, el efecto estimado verdaderamente estima la suma, es decir, $A + BC$. Se dice que A y BC son alias, al igual que B y AC, y que C y AB.

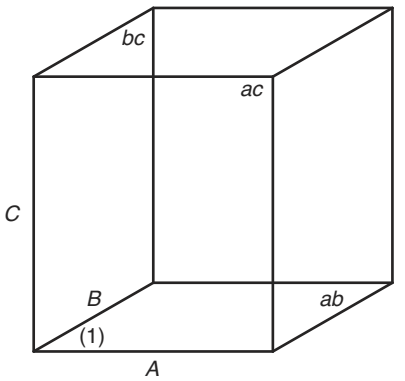
Para la fracción $ABC = -I$ se observa que los alias son los mismos que para la fracción $ABC = I$, además del signo. Así, se tiene

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Las dos fracciones aparecen en las esquinas de los cubos de las figuras 15.15*a* y 15.15*b*.



(a) La fracción $ABC = I$



(a) La fracción $ABC = -I$

Figura 15.15: Las fracciones $\frac{1}{2}$ del factorial 2^3.

Cómo se determinan los alias en general

En general, para un diseño 2^{k-1}, cada efecto, además de aquel definido por el generador, tendrá un *solo socio alias*. El efecto definido por el generador no tendrá alias en otro

efecto, sino que su alias será la media, ya que el estimador de mínimos cuadrados será la media. Para determinar el alias de cada efecto, sólo se comienza con la relación definitoria, digamos $ABC = I$, para el diseño 2^{3-1}. Entonces, para obtener, digamos, el alias para el efecto A, se multiplica A por ambos lados de la ecuación $ABC = I$ y se reduce cualquier exponente por el módulo 2. Por ejemplo,

$$A \cdot ABC = A, \qquad \text{con lo que} \quad BC \equiv A.$$

En forma similar,

$$B \equiv B \cdot ABC \equiv AB^2C \equiv AC,$$

y, por supuesto,

$$C \equiv C \cdot ABC \equiv ABC^2 \equiv AB.$$

Ahora, para la segunda fracción, es decir, la definida por la relación $ABC = -I$,

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Como resultado, el valor numérico del efecto A en realidad estima $A - BC$. De manera similar, el valor de B estima $B - AC$, y el valor de C estima $C - AB$.

Construcción formal del diseño 2^{k-1}

La comprensión plena del concepto de los alias facilita el conocimiento de la construcción del diseño 2^{k-1}. Se comienza con la investigación del 2^{3-1}. Se requieren tres factores y cuatro puntos de diseño. El procedimiento comienza con un **factorial completo** en $k - 1 = 2$ factores A y B. Después se agrega un tercer factor de acuerdo con las estructuras de alias deseadas. Por ejemplo, con ABC como el generador, resulta claro que $C = \pm AB$. Así, se descubre que $C = AB$, o $C = -AB$ complementan el factorial completo en A y B. La tabla 15.10 ilustra un procedimiento que resulta muy sencillo.

Tabla 15.10: Construcción de los dos diseños 2^{3-1}

2^2 básico		$2^{3-1}; ABC = I$			$2^{3-1}; ABC = -I$		
A	B	A	B	$C = AB$	A	B	$C = -AB$
−	−	−	−	+	−	−	−
+	−	+	−	−	+	−	+
−	+	−	+	−	−	+	+
+	+	+	+	+	+	+	−

Note que ya vimos que $ABC = I$ proporciona los puntos de diseño a, b, c y abc; en tanto que $ABC = -I$ proporciona (1), ac, bc y ab. Anteriormente pudimos construir los mismos diseños usando los contrastes que se muestran en la tabla 15.9. Sin embargo, a medida que el diseño se vuelve más complicado con fracciones superiores, esas tablas de contrastes se vuelven más difíciles de trabajar.

Ahora considere un diseño 2^{4-1}, es decir, $\frac{1}{2}$ de un diseño factorial 2^4, que incluye los factores A, B, C y D. Como en el caso del diseño 2^{3-1}, la interacción que se usa como

generador es la interacción de mayor orden, en este caso *ABCD*. Debe recordarse que $ABCD = I$, la relación definitoria sugiere que se sacrifica la información sobre *ABCD*. Aquí comenzamos con el diseño 2^3 completo en *A*, *B* y *C*, y se forma $D = \pm ABC$ para generar los dos diseños 2^{4-1}. La tabla 15.11 ilustra la construcción de ambos diseños.

Tabla 15.11: Construcción de los dos diseños 2^{4-1}

2^3 Básico			2^{4-1}; $ABCD = I$				2^{4-1}; $ABCD = -I$			
A	*B*	*C*	*A*	*B*	*C*	$D = ABC$	*A*	*B*	*C*	$D = -ABC$
−	−	−	−	−	−	−	−	−	−	+
+	−	−	+	−	−	+	+	−	−	−
−	+	−	−	+	−	+	−	+	−	−
+	+	−	+	+	−	−	+	+	−	+
−	−	+	−	−	+	+	−	−	+	−
+	−	+	+	−	+	−	+	−	+	+
−	+	+	−	+	+	−	−	+	+	+
+	+	+	+	+	+	+	+	+	+	−

Aquí, empleando las notaciones *a*, *b*, *c*, etcétera, se tienen los diseños siguientes:

$$ABCD = I, (1), ad, bd, ab, cd, ac, bc, abcd$$
$$ABCD = -I, d, a, b, abd, c, acd, bcd, abc.$$

En el caso de 2^{4-1}, los alias se obtienen como se describió antes para 2^{3-1}. Cada efecto tiene un solo socio alias que se obtiene mediante la multiplicación que se efectúa utilizando la relación definitoria. Por ejemplo, el alias de *A* para el diseño $ABCD = I$ es dado por

$$A = A \cdot ABCD = A^2BCD = BCD.$$

El alias para *AB* es dado por

$$AB = AB \cdot ABCD = A^2B^2CD = CD.$$

Como es fácil observar, los efectos principales tienen alias con interacciones de tres factores y las interacciones de dos factores tienen alias con otras interacciones de dos factores. La lista completa es dada por

$$\begin{aligned} A &= BCD \qquad AB = CD \\ B &= ACD \qquad AC = BD \\ C &= ABD \qquad AD = BC \\ D &= ABC. \end{aligned}$$

Construcción de la fracción de $\frac{1}{4}$

En el caso de la fracción de $\frac{1}{4}$, en vez de una se seleccionan dos interacciones para ser sacrificadas, y la tercera resulta al obtener la interacción generalizada de las dos

seleccionadas. Observe que esto se asemeja mucho a la construcción de cuatro bloques. La fracción que se emplea es simplemente uno de los bloques. Un ejemplo sencillo ayuda mucho a ver la conexión con la construcción de la fracción de $\frac{1}{2}$. Considere la construcción de $\frac{1}{4}$ de un factorial 2^5, es decir, un diseño 2^{5-2} con los factores A, B, C, D y E. Un procedimiento que **evita el confundir dos efectos principales** es la selección de ABD y ACE como las interacciones que corresponden a los dos generadores, lo que produce $ABD = I$ y $ACE = I$ como las relaciones definitorias. La tercera interacción sacrificada sería $(ABD)(ACE) = A^2BCDE = BCDE$. Para la construcción del diseño se comienza con un factorial $2^{5-2} = 2^3$ en A, B y C. Se usan las interacciones ABD y ACE para proporcionar los generadores, de manera que el factorial 2^3 en A, B y C es proporcionado por el factor $D = \pm AB$ y $E = \pm AC$. Así, una de las fracciones es dada por

$$\begin{array}{ccccc|l} A & B & C & D = AB & E = AC & \\ \hline - & - & - & + & + & de \\ + & - & - & - & - & a \\ - & + & - & - & + & be \\ + & + & - & + & - & abd \\ - & - & + & + & - & cd \\ + & - & + & - & + & ace \\ - & + & + & - & - & bc \\ + & + & + & + & + & abcde \end{array}$$

Las otras tres fracciones se calculan utilizando los generadores $\{D = -AB,\ E = AC\}$, $\{D = AB,\ E = -AC\}$ y $\{D = -AB,\ E = -AC\}$. Considere un análisis del diseño 2^{5-2} anterior, que contiene ocho puntos de diseño para estudiar cinco factores. Los alias para los efectos principales son dados por

$$\begin{array}{lll} A(ABD) \equiv BD & A(ACE) \equiv CE & A(BCDE) \equiv ABCDE \\ B \equiv AD & \equiv ABCE & \equiv CDE \\ C \equiv ABCD & \equiv AE & \equiv BDE \\ D \equiv AB & \equiv ACDE & \equiv BCE \\ E \equiv ABDE & \equiv AC & \equiv BCD \end{array}$$

Los alias para otros efectos se pueden obtener de la misma manera. El desglose de los grados de libertad es dado por (además de la réplica)

Efectos principales	5	
Falta de ajuste	2	$(CD = BE, BC = DE)$
Total	7	

Se listan las interacciones sólo para el grado dos en la falta de ajuste.

Ahora considere el caso de un diseño 2^{6-2}, que permite 16 puntos de diseño para estudiar seis factores. Nuevamente se eligen dos generadores de diseño. Una opción pragmática para complementar un factorial $2^{6-2} = 2^4$ completo en A, B, C y D consiste en usar $E = \pm ABC$ y $F = \pm BCD$. La construcción se muestra en la tabla 15.12.

Es evidente que con ocho puntos de diseño más que en 2^{5-2} los alias de los efectos principales no representarán un problema difícil. De hecho, observe que con las relaciones definitorias $ABCE = \pm I$, $BCDF = \pm I$, y $(ABCE)(BCDF) = ADEF = \pm I$, los

Tabla 15.12: Diseño 2^{6-2}

A	B	C	D	$E = ABC$	$F = BCD$	**Combinación de tratamientos**
−	−	−	−	−	−	(1)
+	−	−	−	+	−	*ae*
−	+	−	−	+	+	*bef*
+	+	−	−	−	+	*abf*
−	−	+	−	+	+	*cef*
+	−	+	−	−	+	*acf*
−	+	+	−	−	−	*bc*
+	+	+	−	+	−	*abce*
−	−	−	+	−	+	*df*
+	−	−	+	+	+	*adef*
−	+	−	+	+	−	*bde*
+	+	−	+	−	−	*abd*
−	−	+	+	+	−	*cde*
+	−	+	+	−	−	*acd*
−	+	+	+	−	+	*bcdf*
+	+	+	+	+	+	*abcdef*

efectos principales tendrán alias con interacciones que no son menos complejas que las de tercer orden. La estructura de los alias para los efectos principales se escribe

$$A \equiv BCE \equiv ABCDF \equiv DEF, \qquad D \equiv ABCDE \equiv BCF \equiv AEF,$$
$$B \equiv ACE \equiv CDF \equiv ABDEF, \qquad E \equiv ABC \equiv BCDEF \equiv ADF,$$
$$C \equiv ABE \equiv BDF \equiv ACDEF, \qquad F \equiv ABCEF \equiv BCD \equiv ADE,$$

cada uno con un solo grado de libertad. Para las interacciones de dos factores,

$$AB \equiv CE \equiv ACDF \equiv BDEF, \qquad AF \equiv BCEF \equiv ABCD \equiv DE,$$
$$AC \equiv BE \equiv ABDF \equiv CDEF, \qquad BD \equiv ACDE \equiv CF \equiv ABEF,$$
$$AD \equiv BCDE \equiv ABCF \equiv EF, \qquad BF \equiv ACEF \equiv CD \equiv ABDE,$$
$$AE \equiv BC \equiv ABCDEF \equiv DF.$$

Por supuesto, aquí hay algunos alias entre las interacciones de dos factores. Los dos grados de libertad restantes se explican por medio de los siguientes grupos:

$$ABD \equiv CDE \equiv ACF \equiv BEF, \qquad ACD \equiv BDE \equiv ABF \equiv CEF.$$

Es evidente que antes de recomendar finalmente el plan experimental siempre debemos estar conscientes de que la estructura de alias es para un experimento fraccionario. La selección adecuada de contrastes de definición es importante, ya que es lo que determina la estructura de los alias.

15.7 Análisis de experimentos factoriales fraccionados

La dificultad para realizar pruebas formales de significancia con datos de experimentos factoriales fraccionados radica en la determinación del término del error apropiado.

A menos que se disponga de datos de experimentos anteriores, el error debe provenir de una agrupación de contrastes que representan efectos que se presume son despreciables.

Las sumas de cuadrados para los efectos individuales se calculan usando en esencia los mismos procedimientos que se emplean para obtener el factorial completo. Es posible formar un contraste en las combinaciones de tratamientos construyendo la tabla de signos positivos y negativos. Por ejemplo, para media réplica de un experimento factorial 2^3 con *ABC* como contraste de definición, un conjunto posible de combinaciones de tratamientos, junto con el signo algebraico apropiado para cada contraste que se usa para calcular los efectos y las sumas de cuadrados de los distintos efectos, sería como el que se presenta en la tabla 15.13.

Tabla 15.13: Signos para los contrastes en media réplica de un experimento factorial 2^3

Combinación de tratamientos	**Efecto factorial**						
	A	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+

Observe que en la tabla 15.13 los contrastes *A* y *BC* son idénticos, lo cual ilustra los alias. Asimismo, $B \equiv AC$ y $C \equiv AB$. En esta situación se tienen tres contrastes ortogonales que representan los 3 grados de libertad disponibles. Si se obtuvieran dos observaciones para cada una de las cuatro combinaciones de tratamientos, entonces tendríamos un estimado de la varianza del error con 4 grados de libertad. Si suponemos que los efectos de interacción son despreciables, podríamos probar la significancia de todos los efectos principales.

Un ejemplo del efecto y la suma de cuadrados correspondientes es

$$A = \frac{a - b - c + abc}{2n}, \qquad SCA = \frac{(a - b - c + abc)^2}{2^2 n}.$$

En general, la suma de cuadrados con un grado de libertad para cualquier efecto en una fracción 2^{-p} de un experimento factorial 2^k ($p < k$) se obtiene elevando al cuadrado los contrastes en los totales de los tratamientos seleccionados y dividiendo entre $2^{k-p}n$, donde *n* es el número de réplicas de estas combinaciones de tratamientos.

Ejemplo 15.6: Suponga que se desea emplear una media réplica para estudiar los efectos de cinco factores, cada uno en dos niveles, sobre alguna respuesta, y que se conoce que cualquiera que sea el efecto de cada factor, será constante para cada nivel de los demás factores. En otras palabras, no hay interacciones. Sea el contraste de definición *ABCDE* lo que ocasiona que los efectos principales tengan alias con interacciones de cuatro factores. El agrupamiento de contrastes que incluyen interacciones proporciona $15 - 5 = 10$ grados de libertad para el error. Realice un análisis de varianza con los datos de la tabla 15.14 y pruebe todos los efectos principales a un nivel de significancia de 0.05.

***Solución*:** Las sumas de cuadrados y los efectos para los efectos principales son

$$SCA = \frac{(11.3 - 15.6 - \cdots - 14.7 + 13.2)^2}{2^{5-1}} = \frac{(-17.5)^2}{16} = 19.14,$$

Tabla 15.14: Datos para el ejemplo 15.6

Tratamiento	Respuesta	Tratamiento	Respuesta
a	11.3	*bcd*	14.1
b	15.6	*abe*	14.2
c	12.7	*ace*	11.7
d	10.4	*ade*	9.4
e	9.2	*bce*	16.2
abc	11.0	*bde*	13.9
abd	8.9	*cde*	14.7
acd	9.6	*abcde*	13.2

$$A = -\frac{17.5}{8} = -2.19,$$

$$SCB = \frac{(-11.3 + 15.6 - \cdots - 14.7 + 13.2)^2}{2^{5-1}} = \frac{(18.1)^2}{16} = 20.48,$$

$$B = \frac{18.1}{8} = 2.26,$$

$$SCC = \frac{(-11.3 - 15.6 + \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(10.3)^2}{16} = 6.63,$$

$$C = \frac{10.3}{8} = 1.21,$$

$$SCD = \frac{(-11.3 - 15.6 - \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(-7.7)^2}{16} = 3.71,$$

$$D = \frac{-7.7}{8} = -0.96,$$

$$SCE = \frac{(-11.3 - 15.6 - \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(8.9)^2}{16} = 4.95,$$

$$E = \frac{8.9}{8} = 1.11.$$

Todos los demás cálculos y pruebas de significancia se resumen en la tabla 15.15. Las pruebas indican que el factor *A* tiene un efecto negativo significativo sobre la respuesta; mientras que el factor *B* tiene un efecto positivo significativo. Los factores *C*, *D* y *E* no son significativos al nivel de significancia de 0.05. ∎

Ejercicios

15.18 Liste los alias de los diferentes efectos en un experimento factorial 2^5 cuando el contraste de definición es *ACDE*.

15.19 *a*) Obtenga una fracción de $\frac{1}{2}$ de un diseño factorial 2^4 usando *BCD* como el contraste de definición.

b) Divida la fracción de $\frac{1}{2}$ en dos bloques de cuatro unidades cada uno confundiendo *ABC*.

c) Construya la tabla de análisis de varianza (fuentes de variación y grados de libertad) para probar todos los efectos principales no confundidos, si se acepta que todas las interacciones de los efectos son despreciables.

15.20 Construya una fracción de $\frac{1}{4}$ de un diseño factorial 2^6 utilizando *ABCD* y *BDEF* como los contrastes de definición. Diga cuáles efectos tienen alias con los seis efectos principales.

Tabla 15.15: Análisis de varianza para los datos de media réplica de un experimento factorial 2^5

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Efecto principal:				
A	19.14	1	19.14	6.21
B	20.48	1	20.48	6.65
C	6.63	1	6.63	2.15
D	3.71	1	3.71	1.20
E	4.95	1	4.95	1.61
Error	30.83	10	3.08	
Total	85.74	15		

15.21 *a*) Con los contrastes de definición *ABCE* y *ABDF* obtenga una fracción de $\frac{1}{4}$ de un diseño 2^6.

b) Muestre la tabla del análisis de varianza (fuentes de variación y grados de libertad) para todas las pruebas apropiadas, suponiendo que *E* y *F* no interactúan y que las interacciones de tres factores y mayores son despreciables.

15.22 En un experimento que implica sólo 16 ensayos se varían siete factores en dos niveles. Se utiliza un experimento factorial 2^7 con una fracción de $\frac{1}{8}$, con los contrastes de definición *ACD*, *BEF* y *CEG*. Los datos son los siguientes:

Combinación de tratamientos	Respuesta	Combinación de tratamientos	Respuesta
(1)	31.6	*acg*	31.1
ad	28.7	*cdg*	32.0
abce	33.1	*beg*	32.8
cdef	33.6	*adefg*	35.3
acef	33.7	*efg*	32.4
bcde	34.2	*abdeg*	35.3
abdf	32.5	*bcdfg*	35.6
bf	27.8	*abcfg*	35.1

Realice un análisis de varianza sobre los siete efectos principales, suponiendo que las interacciones son despreciables. Use un nivel de significancia de 0.05.

15.23 Se lleva a cabo un experimento para que un ingeniero adquiera conocimiento acerca de cómo influye la temperatura de sellado *A*, la temperatura de enfriamiento de una barra *B*, el porcentaje de aditivo de polietileno *C* y la presión *D* sobre la resistencia del sello (en gramos por pulgada) de un lote de envoltura para pan. Se utiliza un experimento factorial 2^4 con fracción de $\frac{1}{2}$ con un contraste de definición *ABCD*. A continuación se presentan los datos. Realice un análisis de varianza sólo sobre los efectos principales usando $\alpha = 0.05$

A	*B*	*C*	*D*	Respuesta
−1	−1	−1	−1	6.6
1	−1	−1	1	6.9
−1	1	−1	1	7.9
1	1	−1	−1	6.1
−1	−1	1	1	9.2
1	−1	1	−1	6.8
−1	1	1	−1	10.4
1	1	1	1	7.3

15.24 En un experimento realizado en el Departamento de Ingeniería Mecánica, y analizado por el Centro de Consultoría en Estadística de Virginia Tech, un sensor detecta una carga eléctrica cada vez que las aspas de una turbina completan un giro. Luego, el sensor mide la amplitud de la corriente eléctrica. Seis factores son rpm *A*, temperatura *B*, distancia entre las aspas *C*, distancia entre las aspas y la carcasa *D*, ubicación de la entrada *E*, y ubicación del detector *F*. Se utiliza un experimento factorial 2^6 con fracción de $\frac{1}{4}$, con contrastes de definición *ABCE* y *BCDF*. Los datos son los siguientes:

A	*B*	*C*	*D*	*E*	*F*	Respuesta
−1	−1	−1	−1	−1	−1	3.89
1	−1	−1	−1	1	−1	10.46
−1	1	−1	−1	1	1	25.98
1	1	−1	−1	−1	1	39.88
−1	−1	1	−1	1	1	61.88
1	−1	1	−1	−1	1	3.22
−1	1	1	−1	−1	−1	8.94
1	1	1	−1	1	−1	20.29
−1	−1	−1	1	−1	1	32.07
1	−1	−1	1	1	1	50.76
−1	1	−1	1	1	−1	2.80
1	1	−1	1	−1	−1	8.15
−1	−1	1	1	1	−1	16.80
1	−1	1	1	−1	−1	25.47
−1	1	1	1	−1	1	44.44
1	1	1	1	1	1	2.45

Realice un análisis de varianza de los efectos principales y las interacciones de dos factores, si se acepta que las interacciones de tres factores o más son despreciables. Use $\alpha = 0.05$.

15.25 En el estudio denominado *Durability of Rubber to Steel Adhesively Bonded Joints*, efectuado por el Departamento de Ciencias del Ambiente y Mecánica, y analizado por el Centro de Consultoría en Estadística de Virginia Tech, un experimentador midió el número de roturas en un sello adhesivo. Se planteó que la concentración de agua marina *A*, la temperatura *B*, el pH *C*, el voltaje *D* y la tensión *E* influyen en el rompimiento de un sello adhesivo. Se utilizó un experimento factorial 2^5 con fracción de $\frac{1}{2}$ y con el contraste de definición *ABCDE*. Los datos son los siguientes:

A	*B*	*C*	*D*	*E*	Respuesta
−1	−1	−1	−1	1	462
1	−1	−1	−1	−1	746
−1	1	−1	−1	−1	714
1	1	−1	−1	1	1070
−1	−1	1	−1	−1	474
1	−1	1	−1	1	832
−1	1	1	−1	1	764
1	1	1	−1	−1	1087
−1	−1	−1	1	−1	522
1	−1	−1	1	1	854
−1	1	−1	1	1	773
1	1	−1	1	−1	1068
−1	−1	1	1	1	572
1	−1	1	1	−1	831
−1	1	1	1	−1	819
1	1	1	1	1	1104

Realice un análisis de varianza de los efectos principales y de las interacciones de dos factores *AD, AE, BD, BE*; suponga que las interacciones de tres o más factores son despreciables. Use $\alpha = 0.05$.

15.26 Considere un diseño 2^{5-1} con los factores *A*, *B*, *C*, *D* y *E*. Construya el diseño comenzando con un diseño 2^4 y use $E = ABCD$ como generador. Indique todos los alias.

15.27 Hay seis factores y sólo se pueden usar ocho puntos de diseño. Construya un diseño 2^{6-3}, comenzando con un diseño 2^3, y utilice $D = AB$, $E = -AC$ y $F = BC$ como generadores.

15.28 Considere el ejercicio 15.27. Construya otro 2^{6-3} que sea diferente del diseño elegido en el ejercicio 15.27.

15.29 Para el ejercicio 15.27 proporcione todos los alias para los seis efectos principales.

15.30 En Myers, Montgomery y Anderson-Cook (2009) se analiza una aplicación en la cual a un ingeniero le interesan los efectos del agrietamiento de una aleación de titanio. Los tres factores son *A*, temperatura; *B*, contenido de titanio; y *C*, cantidad de refinador en grano. La siguiente tabla presenta una parte del diseño y la respuesta, la longitud de las grietas inducida en la muestra de la aleación.

A	*B*	*C*	Respuesta
−1	−1	−1	0.5269
1	1	−1	2.3380
1	−1	1	4.0060
1	1	1	3.3640

a) ¿Cuál es la relación de definición?
b) Proporcione alias para los tres efectos principales asumiendo que las interacciones de dos factores pueden ser reales.
c) Si suponemos que las interacciones son despreciables, ¿cuál será el factor principal más importante?
d) ¿Qué nivel sugeriría para el factor obtenido en el inciso *c* en la producción final, alto o bajo?
e) ¿Qué niveles sugeriría para los demás factores en la producción final?
f) ¿Qué riesgos hay en las recomendaciones que hizo en los incisos *d* y *e*? Responda de manera detallada.

15.8 Diseños de fracciones superiores y de filtrado

Algunas situaciones industriales requieren que el analista determine cuáles factores controlables, de entre un número grande de ellos, tienen un efecto sobre alguna respuesta importante. Los factores pueden ser cualitativos o variables de clase, variables de regresión o una mezcla de ambas. El procedimiento analítico puede requerir un análisis de varianza, una regresión o ambos. A menudo el modelo de regresión utilizado sólo incluye los efectos lineales principales, aunque tal vez sea posible estimar algunas interacciones. La situación exige la selección de variables y los diseños experimentales resultantes se denominan **diseños de filtrado**. Es evidente que los diseños ortogonales de dos niveles saturados o casi saturados son candidatos viables.

Resolución del diseño

A menudo los diseños ortogonales de dos niveles se clasifican según su **resolución**, la cual es determinada por la siguiente definición.

Definición 15.1: La **resolución** de un diseño ortogonal de dos niveles es la longitud de la interacción más pequeña (menos compleja) de entre el conjunto de contrastes de definición.

Si el diseño se construye como un factorial completo o fraccionado, ya sea un diseño 2^k, o bien, 2^{k-p}, $p = 1, 2, \ldots, k-1$, el concepto de resolución del diseño es un auxiliar para determinar el efecto de los alias. Por ejemplo, un diseño de resolución II sería de poca utilidad, ya que habría al menos un caso de alias de un efecto principal con otro. Un diseño de resolución III tendría todos sus efectos principales (lineales) ortogonales entre sí. No obstante, habrá algunos alias entre los efectos lineales y las interacciones de dos factores. Entonces, es evidente que si el analista está interesado en estudiar los efectos principales (lineales en el caso de la regresión) y no hay interacciones de dos factores, entonces se requiere un diseño cuya resolución sea de al menos III.

15.9 Construcción de diseños de resolución III y IV, con 8, 16 y 32 puntos de diseño

Es posible construir diseños útiles con resoluciones III y IV para 2 a 7 variables con 8 puntos de diseño. Empezamos con un factorial 2^3 que haya sido saturado simbólicamente con interacciones.

$$\begin{array}{c} x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_2x_3 \quad x_1x_2x_3 \\ \left[\begin{array}{rrrrrrr} -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right] \end{array}$$

Es evidente que, con sólo reemplazar las columnas de interacción por nuevos efectos principales para las siete variables, se puede construir un diseño de resolución III. Por ejemplo, podríamos definir

$$\begin{array}{ll} x_4 = x_1x_2 & \text{(contraste de definición } ABD\text{)} \\ x_5 = x_1x_3 & \text{(contraste de definición } ACE\text{)} \\ x_6 = x_2x_3 & \text{(contraste de definición } BCF\text{)} \\ x_7 = x_1x_2x_3 & \text{(contraste de definición } ABCG\text{)} \end{array}$$

y obtendríamos una fracción 2^{-4} de un factorial 2^7. Las expresiones anteriores identifican los contrastes de definición elegidos. Resultan once contrastes de definición adicionales y todos contienen al menos tres letras. Así, el diseño es de resolución III. Es evidente que si se comienza con un *subconjunto* de columnas aumentadas y se concluye con un diseño

Tabla 15.16: Algunos diseños 2^{k-p} de resoluciones III, IV, V, VI y VII

Número de factores	Diseño	Número de puntos	Generadores
3	2^{3-1}_{III}	4	$C = \pm AB$
4	2^{4-1}_{IV}	8	$D = \pm ABC$
5	2^{5-2}_{III}	8	$D = \pm AB\ ;\ E = \pm AC$
6	2^{6-1}_{VI}	32	$F = \pm ABCDE$
	2^{6-2}_{IV}	16	$E = \pm ABC\ ;\ F = \pm BCD$
	2^{6-3}_{III}	8	$D = \pm AB\ ;\ F = \pm BC\ ;\ E = \pm AC$
7	2^{7-1}_{VII}	64	$G = \pm ABCDEF$
	2^{7-2}_{IV}	32	$E = \pm ABC\ ;\ G = \pm ABDE$
	2^{7-3}_{IV}	16	$E = \pm ABC\ ;\ F = \pm BCD\ ;\ G = \pm ACD$
	2^{7-4}_{III}	8	$D = \pm AB\ ;\ E = \pm AC\ ;\ F = \pm BC\ ;\ G = \pm ABC$
8	2^{8-2}_{V}	64	$G = \pm ABCD\ ;\ H = \pm ABEF$
	2^{8-3}_{IV}	32	$F = \pm ABC\ ;\ G = \pm ABD\ ;\ H = \pm BCDE$
	2^{8-4}_{IV}	16	$E = \pm BCD\ ;\ F = \pm ACD\ ;\ G = \pm ABC\ ;\ H = \pm ABD$

que incluye menos de 7 variables de diseño, el resultado es un diseño de resolución III en menos de siete variables.

Es posible construir un conjunto similar de diseños posibles para 16 puntos de diseño, comenzando con un diseño 2^4 saturado con interacciones. Las definiciones de las variables que corresponden a estas interacciones producen diseños de resolución III por medio de 15 variables. De manera similar, se pueden construir diseños que contengan 32 corridas, comenzando con un diseño 2^5.

La tabla 15.16 proporciona lineamientos para construir diseños de 8, 16, 32 y 64 puntos, con resolución III, IV e incluso V. La tabla proporciona el número de factores, el número de corridas y los generadores que se utilizan para producir los diseños 2^{k-p}. El generador dado se emplea para **aumentar el factorial completo** que contiene $k - p$ factores.

15.10 Otros diseños de resolución III de dos niveles; los diseños de Plackett-Burman

Una familia de diseños desarrollada por Plackett y Burman (1946, véase la bibliografía) llena el vacío del tamaño de la muestra que existe con los factoriales fraccionados. Éstos son útiles con muestras de tamaño 2^r, es decir, incluyen muestras de tamaños 4, 8, 16, 32, 64,... Los diseños de Plackett -Burman incluyen $4r$ puntos de diseño, por lo que se dispone de diseños de tamaño 12, 20, 24, 28, etcétera. Estos diseños de Plackett-Burman de dos niveles son diseños de resolución III y son muy fáciles de construir. Se proporcionan "renglones básicos" para cada tamaño de muestra. Estos renglones de signos + y − son $n - 1$ en número. Para construir las columnas de la matriz de diseño se comienza con el renglón básico y se hace una permutación cíclica sobre las columnas, hasta que se forman k columnas (el número deseado de variables). Después se llena el último

renglón con signos negativos. El resultado será un diseño de resolución III con k variables ($k = 1, 2, \ldots, N$). Los renglones básicos son los siguientes:

$N = 12$	+ + − + + + − − − + −
$N = 16$	+ + + + − + − + + − − + − − −
$N = 20$	+ + − − + + + + − + − + − − − − + + −
$N = 24$	+ + + + + − + − + + − − + + − − + − + − − − −

Ejemplo 15.7 Construya un diseño depurado de dos niveles con 6 variables que contengan 12 puntos de diseño.

***Solución*:** Comience con el renglón básico en la columna inicial. La segunda columna se forma llevando la entrada inferior de la primera columna a la parte superior de la segunda, y repitiendo la primera. La tercera columna se forma del mismo modo, utilizando las entradas de la segunda columna. Cuando haya un número suficiente de columnas **sencillamente se llena el último renglón con signos negativos**. El diseño resultante es como sigue:

x_1	x_2	x_3	x_4	x_5	x_6
+	−	+	−	−	−
+	+	−	+	−	−
−	+	+	−	+	−
+	−	+	+	−	+
+	+	−	+	+	−
+	+	+	−	+	+
−	+	+	+	−	+
−	−	+	+	+	−
−	−	−	+	+	+
+	−	−	−	+	+
−	+	−	−	−	+
−	−	−	−	−	−

Los diseños de Plackett-Burman son populares en la industria para situaciones de filtrado. Como se trata de diseños de resolución III. todos los efectos lineales son ortogonales. Para cualquier tamaño de muestra el usuario dispone de un diseño para $k = 2, 3, \ldots, N - 1$ variables.

La estructura de alias para el diseño de Plackett-Burman es muy complicada, por lo que el usuario no puede construir el diseño con un control completo de la estructura de alias, como en el caso de los diseños 2^k o 2^{k-p}. Sin embargo, en el caso de modelos de regresión el diseño de Plackett-Burman acepta interacciones (aunque no serán ortogonales) cuando se dispone de suficientes grados de libertad. ■

15.11 Introducción a la metodología de superficie de respuesta

En el estudio de caso 15.2 se ajustó un modelo de regresión a un conjunto de datos con la meta específica de encontrar condiciones en esas variables de diseño que optimizaran (maximizaran) la eficiencia de purificación del carbón. El modelo incluía tres efectos principales lineales, tres términos de interacción de dos factores y un término de interacción de tres factores. La respuesta del modelo era la eficiencia de la purificación, y las condiciones óptimas de x_1, x_2 y x_3 se obtuvieron utilizando los signos y la magnitud

de los coeficientes del modelo. En este ejemplo se utilizó un diseño de dos niveles para mejorar el proceso o para optimizarlo. En muchas áreas de la ciencia y de la ingeniería la aplicación se extiende para incluir modelos y diseños más complicados a los que, en conjunto, se les denomina **metodología de superficie de respuesta (MSR)**. Esta metodología abarca tanto métodos gráficos como analíticos. El término *superficie de respuesta* se deriva de la apariencia de la superficie multidimensional de la respuesta estimada constante de un modelo de segundo orden, es decir, un modelo con términos de primer y segundo orden. A continuación se presenta un ejemplo.

El modelo de superficie de respuesta de segundo orden

En muchos ejemplos industriales de optimización de procesos se utiliza un *modelo de superficie de respuesta de segundo orden*. Para el caso de, digamos $k = 2$ variables de proceso o variables de diseño, y una sola respuesta y, el modelo es dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon.$$

Aquí se tienen $k = 2$ términos de primer orden, dos términos puros de segundo orden o cuadráticos y un término de interacción dado por $\beta_{12} x_1 x_2$. Los términos x_1 y x_2 se codifican en la forma conocida de ± 1. El término ϵ denota al acostumbrado error del modelo. En general, para k variables de diseño el modelo contendrá $1 + k + k + \binom{k}{2}$ términos del modelo y, por lo tanto, el diseño experimental debe contener al menos un número similar de puntos de diseño. Además, los términos cuadráticos requieren que las variables de diseño estén fijas en el diseño con al menos tres niveles. Al diseño resultante se le denomina *diseño de segundo orden*. A continuación se presenta un ejemplo.

El siguiente **diseño central compuesto (DCC)** y el ejemplo fueron tomados de Myers, Montgomery y Anderson-Cook (2009). Quizás la clase más popular de diseños de segundo orden sea la clase de los diseños centrales compuestos. El ejemplo que se presenta en la tabla 15.17 se refiere a un proceso químico en el que la temperatura de reacción, ξ_1, y la concentración del reactante, ξ_2, se muestran en sus niveles naturales y también de forma codificada. Cada factor tiene cinco niveles. Además, se incluye el orden en que se realizaron las observaciones de x_1 y x_2. La columna de la derecha proporciona los valores de la respuesta y, el porcentaje de conversión del proceso. Los primeros cuatro puntos de diseño representan los conocidos puntos factoriales en los niveles ± 1. Los siguientes cuatro puntos se conocen como puntos axiales, los cuales van seguidos por las corridas centrales que se explicaron y ejemplificaron antes en este capítulo. De esta manera, los cinco niveles de cada uno de los dos factores son -1, $+1$, -1.414, $+1.414$ y 0. En la figura 15.16 se presenta una imagen clara de la geometría del diseño central compuesto para este ejemplo de $k = 2$. En esta figura se ilustra la fuente del término **puntos axiales**. Estos cuatro puntos se localizan sobre los ejes factoriales, a una distancia axial de $\alpha = \sqrt{2} = 1.414$ a partir del centro del diseño. De hecho, para este DCC en particular, los puntos del perímetro, axiales y factoriales, se encuentran todos a la distancia $\sqrt{2}$ del centro del diseño, y como resultado tenemos ocho puntos equidistantes sobre un círculo más cuatro réplicas en el centro del diseño.

Ejemplo 15.8: **Análisis de superficie de respuesta:** Un análisis de los datos en el ejemplo de las dos variables podría implicar el ajuste de una función de respuesta de segundo orden. La superficie de respuesta resultante se puede utilizar de forma analítica o gráfica para determinar el impacto que tienen x_1 y x_2 sobre el porcentaje de conversión del proceso. Los coeficientes en la función de respuesta están determinados por medio del método de

Tabla 15.17: Diseño central compuesto para el ejemplo 15.8

		Temperatura (°C)	Concentración (%)			
Observación	**Corrida**	ξ_1	ξ_2	x_1	x_2	y
1	4	200	15	−1	−1	43
2	12	250	15	1	−1	78
3	11	200	25	−1	1	69
4	5	250	25	1	1	73
5	6	189.65	20	−1.414	0	48
6	7	260.35	20	1.414	0	78
7	1	225	12.93	0	−1.414	65
8	3	225	27.07	0	1.414	74
9	8	225	20	0	0	76
10	10	225	20	0	0	79
11	9	225	20	0	0	83
12	2	225	20	0	0	81

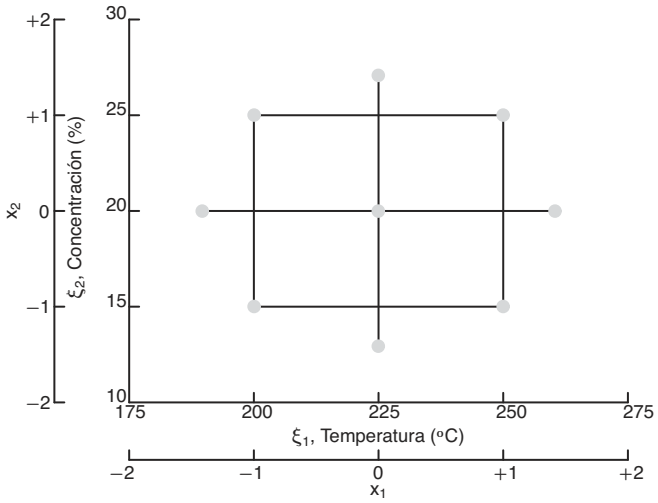


Figura 15.16: Diseño central compuesto para el ejemplo 15.8.

mínimos cuadrados que estudiamos en el capítulo 12 y que ejemplificamos a lo largo de este capítulo. El modelo resultante de respuesta de segundo orden es dado en las variables codificadas como

$$\hat{y} = 79.75 + 10.18x_1 + 4.22x_2 - 8.50x_1^2 - 5.25x_2^2 - 7.75x_1x_2,$$

mientras que en las variables naturales es dado por

$$\hat{y} = -1080.22 + 7.7671\xi_1 + 23.1932\xi_2 - 0.0136\xi_1^2 - 0.2100\xi_2^2 - 0.0620\xi_1\xi_2.$$

Como este ejemplo sólo incluye dos variables de diseño, el método más esclarecedor para determinar la naturaleza de la superficie de respuesta en la región del diseño

consiste en utilizar gráficas de dos o tres dimensiones. Sería interesante determinar cuáles niveles de temperatura x_1 y concentración x_2 producen un estimado deseable del porcentaje de conversión $\hat{y}$. La función de respuesta estimada anterior se graficó en tres dimensiones y la *superficie de respuesta* resultante se presenta en la figura 15.17. La altura de la superficie es $\hat{y}$ expresada en porcentaje. En esta figura es fácil observar por qué se utiliza el término **superficie de respuesta**. En el caso en que sólo se utilizan dos variables de diseño las gráficas bidimensionales de curvas pueden ser útiles. Observe en la figura 15.18 que las curvas de la conversión constante estimada se ven como rodajas de la superficie de respuesta. Observe que cualquiera de las dos figuras indica con facilidad cuáles coordenadas de temperatura y concentración producen el mayor porcentaje de conversión estimado. En las gráficas las coordenadas se presentan tanto en unidades codificadas como en unidades naturales. Observe que la mayor conversión estimada se encuentra en aproximadamente 240°C y una concentración de 20%. La respuesta máxima estimada (o pronosticada) en esa ubicación es 82.47%.

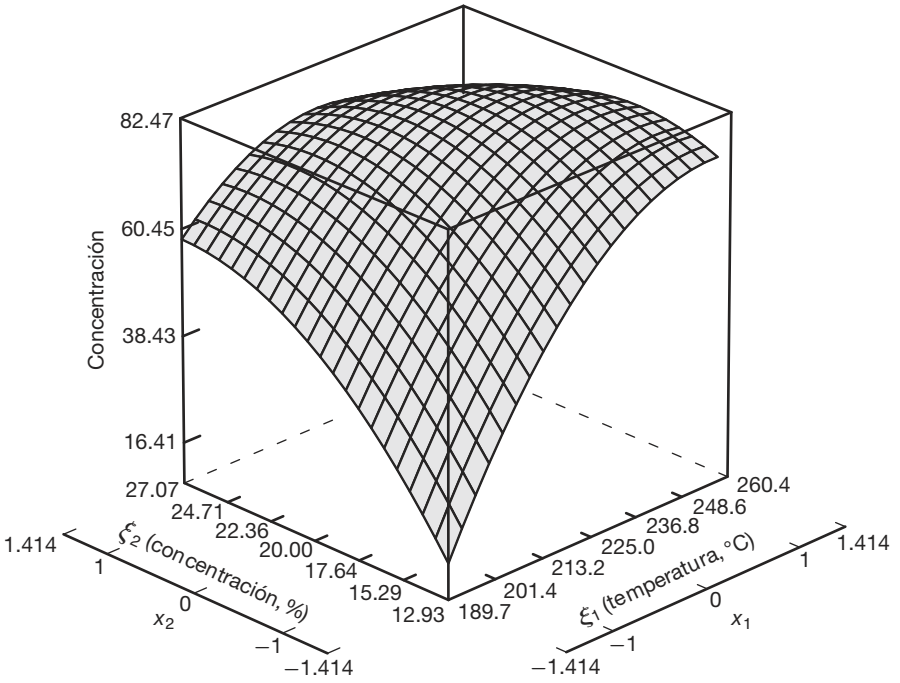


Figura 15.17: Gráfica de la superficie de respuesta de la conversión pronosticada para el ejemplo 15.8.

Otros comentarios acerca del análisis de superficie de respuesta

El libro de Myers, Montgomery y Anderson-Cook (2009) proporciona una gran cantidad de información sobre el análisis y el diseño de la metodología de superficie de respuesta. La ilustración gráfica que se utilizó aquí podría ampliarse con resultados analíticos que brindan información acerca de la naturaleza de la superficie de respuesta dentro de la región del diseño.

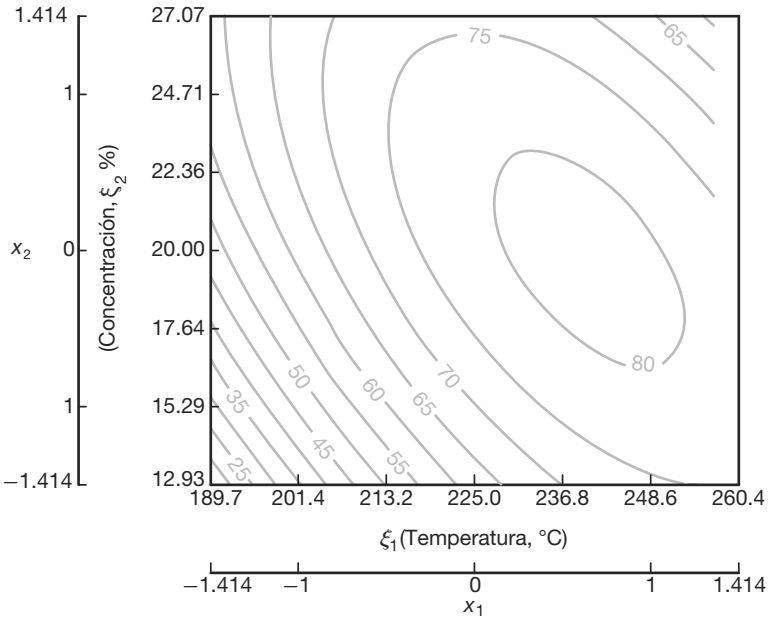


Figura 15.18: Gráfica de curvas de la conversión pronosticada para el ejemplo 15.8.

Se pueden usar otros cálculos para determinar si la ubicación de las condiciones óptimas está dentro o muy lejos de la región del diseño experimental. Existen muchos aspectos importantes a tomar en cuenta cuando se necesita determinar las condiciones apropiadas para la operación futura de un proceso.

Otras secciones del libro de Myers, Montgomery y Anderson-Cook (2009) abordan otros aspectos del diseño experimental. Por ejemplo, el diseño central compuesto, aunque es el tipo de diseño más útil, no es el único que se utiliza en la metodología de superficie de respuesta. En el libro mencionado se analizan muchos otros tipos. Además, el diseño central compuesto al que aquí nos referimos es un caso especial en el que $k = 2$. El caso más general $k > 2$ se analiza en Myers, Montgomery y Anderson-Cook (2009).

15.12 Diseño robusto de parámetros

En este capítulo se destacó el concepto del uso del diseño de experimentos (DE) para adquirir conocimientos sobre procesos de ingeniería y científicos. En el caso en que un proceso incluye un producto es posible usar el DE para mejorar el producto o la calidad. Como se expuso en el capítulo 1, se ha dado mucha importancia al empleo de métodos estadísticos para mejorar los productos. Un aspecto importante de este esfuerzo por mejorar la calidad, que surgió en la década de 1980 y continuó a lo largo de la década de 1990, consiste en incluir la calidad en los procesos y productos en la etapa de investigación o de diseño del proceso. A menudo se requiere del DE para desarrollar procesos con las siguientes propiedades:

1. Insensibles (robustos) a las condiciones ambientales

2. Insensibles (robustos) a factores que dificultan el control
3. Que proporcionen la mínima variación de desempeño

Los métodos que se utilizan para lograr las características deseables en los puntos 1, 2 y 3 forman parte de lo que se conoce como *diseño robusto de parámetros* o DRP (véase Taguchi, 1991; Taguchi y Wu, 1985; y Kackar, 1985, en la bibliografía). En este contexto el término *diseño* se refiere al diseño de los procesos o sistemas, en tanto que *parámetro* se refiere a los parámetros en el sistema. Éstos son a los que nos hemos referido como *factores* o *variables*.

Queda muy claro que las metas 1, 2 y 3 mencionadas son muy nobles. Por ejemplo, un ingeniero petrolero puede tener una buena mezcla de gasolina que se desempeñe muy bien en condiciones ideales y estables. Sin embargo, el desempeño se deteriorará debido a cambios en las condiciones ambientales, como tipo de conductor, factores climáticos, tipo de motor, etc. Un científico de una empresa de alimentos podría tener una muy buena mezcla para pasteles, a menos que el usuario no siga con exactitud las instrucciones del empaque con respecto a la temperatura del horno, tiempo de horneado, entre otros. Un producto o proceso cuyo desempeño sea consistente cuando se expone a esas condiciones ambientales cambiantes se denomina **producto robusto** o **proceso robusto**. (Véase Myers, Montgomery y Anderson-Cook, 2009, en la bibliografía).

Variables de control y ruido

Taguchi (1991) destacó la idea de utilizar dos clases de variables de diseño en un estudio que incluye un diseño de superficie de respuesta (DSR): *factores de control* y *factores de ruido.*

Definición 15.2: Los **factores de control** son variables que se pueden controlar tanto en el experimento como en el proceso. Los **factores de ruido** son variables que pueden o no controlarse en el experimento, pero que no pueden controlarse en el proceso (o que no pueden controlarse bien).

Un método importante consiste en usar variables de control y variables de ruido en el mismo experimento, como efectos fijos. Para lograr esto con frecuencia se utilizan los diseños o arreglos ortogonales.

Meta del diseño robusto de parámetros

La meta del diseño robusto de parámetros es elegir los niveles de las variables de control, es decir, el diseño del proceso, que sean más robustos (insensibles) a los cambios en las variables de ruido.

Debe señalarse que los *cambios en las variables de ruido* en realidad implican cambios durante el proceso, cambios en el campo, cambios en el ambiente, cambios en el manejo o uso por parte del consumidor, etcétera.

Arreglo del producto

Un enfoque del diseño de experimentos que incluye tanto variables de control como de ruido consiste en utilizar un plan experimental que requiere un diseño ortogonal para las variables de control y de ruido, por separado. Entonces, el experimento completo es simplemente el producto o cruce de estos dos diseños ortogonales. El siguiente es un ejemplo sencillo de un arreglo de productos con dos variables de control y dos de ruido.

Ejemplo 15.9: En el artículo "The Taguchi Approach to Parameter Design" en *Quality Progress*, de diciembre de 1987, D. M. Byrne y S. Taguchi analizan un ejemplo interesante en el que se busca un método para ensamblar un conector electrométrico a un tubo de nailon que entrega el rendimiento de arranque requerido para una aplicación de motor automotriz. El objetivo es encontrar condiciones controlables que maximicen la fuerza de arranque. Entre las variables controlables están A, el espesor de la pared del conector, y B, la profundidad de inserción. Durante la operación rutinaria existen diversas variables que no se pueden controlar, aunque se controlan durante el experimento. Entre ellas están C, el tiempo de acondicionamiento, y D, la temperatura de acondicionamiento. Se toman tres niveles para cada variable de control y dos para cada variable de ruido. Como resultado, el arreglo cruzado es el siguiente. Se trata de un arreglo de control de 3×3 y el de ruido es el conocido factorial 2^2 con (1), c, d y cd que representan las combinaciones de los cuatro factores. El propósito del factor de ruido es crear la *clase de variabilidad de la respuesta, la fuerza de arranque, que se podría esperar en la operación diaria con el proceso.* En la tabla 15.18 se muestra el diseño. ▟

Tabla 15.18: Diseño para el ejemplo 15.9

		***B* (profundidad)**		
		Superficial	**Media**	**Profunda**
***A* (espesor de pared)**	**Delgado**	(1)	(1)	(1)
		c	*c*	*c*
		d	*d*	*d*
		cd	*cd*	*cd*
	Medio	(1)	(1)	(1)
		c	*c*	*c*
		d	*d*	*d*
		cd	*cd*	*cd*
	Grueso	(1)	(1)	(1)
		c	*c*	*c*
		d	*d*	*d*
		cd	*cd*	*cd*

Estudio de caso 15.3: **Optimización de proceso de soldadura.** En un experimento que Schmidt y Launsby describen en *Understanding Industrial Designed Experiments* (1991; véase la bibliografía), en una planta de ensamble de circuitos integrados se lleva a cabo la optimización de un proceso de soldadura. Las partes se insertan a mano o en forma automática en una tarjeta que tiene impreso un circuito. Una vez que las partes se insertan, la tarjeta se coloca en una máquina soldadora de ola que se emplea para conectar todos los elementos del circuito. Las tarjetas se colocan en un transportador y pasan por una serie de etapas. Se lavan en una mezcla fundente para eliminar el óxido. Para minimizar la torsión se precalientan antes de aplicar la soldadura, la cual se realiza conforme las tarjetas se mueven a través de la ola de soldadura. El objetivo del experimento consiste en minimizar el número de defectos de soldadura por millón de uniones. Los factores y los niveles de control se incluyen en la tabla 15.19.

Tabla 15.19: Factores de control para el estudio de caso 15.3

Factor	(-1)	$(+1)$
A, temperatura del crisol para soldar (°F)	480	510
B, velocidad del transportador (pies/min)	7.2	10
C, densidad del fundente	0.9°	1.0°
D, temperatura de precalentado	150	200
E, altura de la ola (pulgadas)	0.5	0.6

A nivel experimental es fácil controlar estos factores, pero en la planta o en el proceso es mucho más difícil.

Factores de ruido: tolerancias sobre los factores de control

A menudo, en procesos como éste los factores naturales de ruido son las tolerancias sobre los factores de control. Por ejemplo, en el proceso real en línea la temperatura del crisol para soldar y la velocidad de la banda transportadora son difíciles de controlar. Se sabe que el control de la temperatura está dentro de ± 5°F, y que el control de la velocidad de la banda está dentro de ± 0.2 pies/min. Es posible que la variabilidad de la respuesta del producto (desempeño de la soldadura) se incremente debido a la incapacidad de controlar esos dos factores en ciertos niveles nominales. El tercer factor de ruido es el tipo de ensamble involucrado. En la práctica se utilizan uno de dos tipos de ensambles. Así, se tienen los factores de ruido que se presentan en la tabla 15.20.

Tabla 15.20: Factores de ruido para el estudio de caso 15.3

Factor	(-1)	$(+1)$
*A**, tolerancia de la temperatura del crisol para soldar (°F), (desviación de la nominal)	-5	$+5$
*B**, tolerancia de la velocidad del transportador (pies/min), (desviación del ideal)	-0.2	$+0.2$
*C**, tipo de ensamble	1	2

Se eligieron factoriales fraccionados tanto para el arreglo de control (arreglo interior) como para el de ruido (arreglo exterior): el primero es $\frac{1}{4}$ de un diseño 2^5, y el segundo es $\frac{1}{2}$ de un diseño 2^3. El arreglo cruzado y los valores de respuesta se presentan en la tabla 15.21. Las primeras tres columnas del arreglo interior representan un diseño 2^3. La cuarta y la quinta columnas están formadas por $D = -AC$ y $E = -BC$. Así, las interacciones de definición para el arreglo interior son *ACD*, *BCE* y *ABDE*. El arreglo exterior es una fracción estándar de resolución III de un diseño 2^3. Observe que cada punto del arreglo interior contiene corridas del arreglo exterior. Así, se observan cuatro valores de respuesta en cada combinación del arreglo de control. La figura 15.19 muestra gráficas que revelan el efecto de la temperatura y la densidad sobre la respuesta media.

Tabla 15.21: Arreglos cruzados y valores de respuesta para el estudio de caso 15.3

Arreglo interior					Arreglo exterior					
A	***B***	***C***	***D***	***E***	**(1)**	***a*b****	***a*c****	***b*c****	$\bar{y}$	s_y
1	1	1	−1	−1	194	197	193	275	214.75	40.20
1	1	−1	1	1	136	136	132	136	135.00	2.00
1	−1	1	−1	1	185	261	264	264	243.50	39.03
1	−1	−1	1	−1	47	125	127	42	85.25	47.11
−1	1	1	1	−1	295	216	204	293	252.00	48.75
−1	1	−1	−1	1	234	159	231	157	195.25	43.04
−1	−1	1	1	1	328	326	247	322	305.75	39.25
−1	−1	−1	−1	−1	186	187	105	104	145.50	47.35

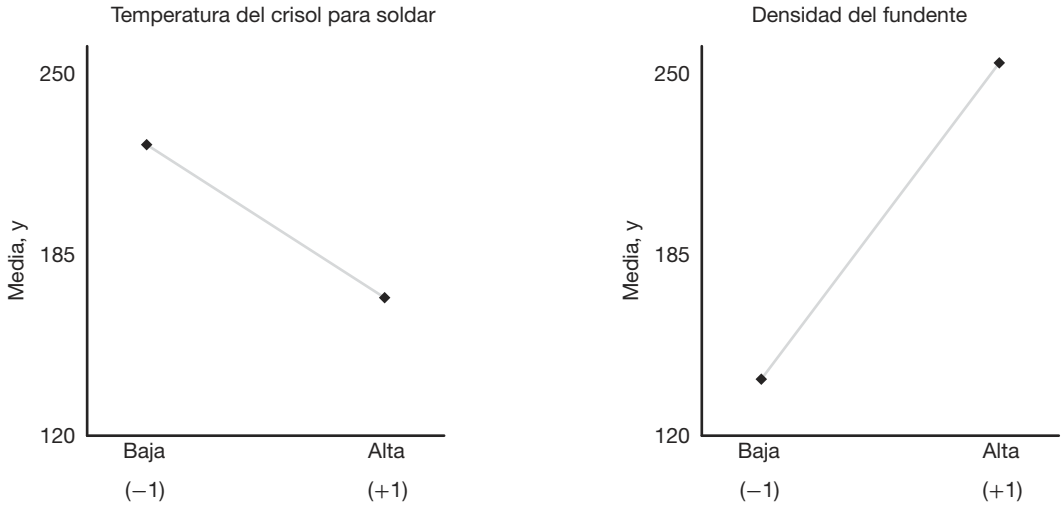


Figura 15.19: Gráfica que muestra la influencia de los factores sobre la respuesta media.

Análisis simultáneo de la media y varianza del proceso

En la mayoría de los ejemplos que utilizan DSR el analista se interesa por encontrar condiciones para las variables de control que proporcionen valores adecuados para la respuesta media $\bar{y}$. Sin embargo, la variación de las variables de ruido proporciona información acerca de la varianza del proceso σ_y^2 que podría anticiparse en el mismo. Es evidente que un producto robusto es aquel para el que el proceso es consistente y, por lo tanto, tiene poca varianza. El DSR puede incluir el análisis simultáneo de $\bar{y}$ y s_y.

Resulta que la temperatura y la densidad del fundente son los factores más importantes en el estudio de caso 15.3, y al parecer influyen en s_y y $\bar{y}$. Por fortuna, para ambas es preferible una *alta temperatura* y una *baja densidad del fundente*. De acuerdo con la figura 15.19 las condiciones "óptimas" son

$$\text{temperatura de soldadura} = 510°\text{F}, \qquad \text{densidad del fundente} = 0.9°.$$

Enfoques alternativos al diseño robusto de parámetros

Un enfoque sugerido por muchos estudiosos consiste en modelar la media y la varianza muestrales por separado. Con frecuencia el modelado separado ayuda al experimentador a comprender mejor el proceso involucrado. En el siguiente ejemplo se ilustra este enfoque con el experimento del proceso de soldadura.

Estudio de caso 15.4: Considere los datos del estudio de caso 15.3. Un método alternativo consiste en ajustar modelos separados para la media $\bar{y}$ y la desviación estándar muestral. Suponga que se usa el código habitual $+1$ y -1 para los factores de control. Con base en la importancia aparente de la temperatura del crisol para soldar x_1 y la densidad del fundente x_2, la regresión lineal sobre la respuesta (número de errores por millón de uniones) produce

$$\hat{y} = 197.125 - 27.5x_1 + 57.875x_2.$$

Para obtener los niveles más robustos de la temperatura y la densidad del fundente es conveniente establecer un compromiso entre la respuesta media y la variabilidad, y para esto es necesario modelar la variabilidad. Una herramienta importante para hacerlo es la transformación logarítmica (véase Bartlett y Kendall, 1946, o Carroll y Ruppert, 1988):

$$\text{In } s^2 = \gamma_0 + \gamma_1(x_1) + \gamma_2(x_2).$$

Este proceso de modelado produce el siguiente resultado:

$$\widehat{\ln s^2} = 6.6975 - 0.7458x_1 + 0.6150x_2.$$

El modelo *logarítmico lineal* tiene un amplio uso en el modelado de la varianza muestral, ya que la transformación logarítmica de la varianza muestral se presta al uso del método de mínimos cuadrados. Esto resulta del hecho de que las suposiciones de normalidad y de varianza homogénea a menudo son muy buenas cuando se utiliza $\ln s^2$ en lugar de s^2 como respuesta del modelo.

El análisis que es importante para el científico o el ingeniero echa mano de los dos modelos al mismo tiempo. Un método gráfico puede ser muy útil. La figura 15.20 presenta al mismo tiempo gráficas sencillas de los modelos de la media y de la desviación estándar. Como se esperaría, la ubicación de la temperatura y la densidad del fundente que minimizan el número medio de errores es la misma que la que minimiza la variabilidad, es decir, temperatura alta y densidad del flujo baja. El método gráfico de la *superficie múltiple de respuesta* permite que el usuario perciba intercambios entre la media del proceso y su variabilidad. Para este ejemplo es probable que el ingeniero se sienta insatisfecho con las condiciones extremas de la temperatura de la soldadura y la densidad del fundente. La figura ofrece estimados de lo que se pierde a medida que uno se aleja de las condiciones óptimas de la media y la variabilidad hacia cualquier condición intermedia. ◼

En el estudio de caso 15.4 para las variables de control se eligieron valores que proporcionaran condiciones deseables tanto para la media como para la varianza del proceso. Se tomaron la media y la varianza a través de la distribución de las variables de ruido en el proceso y se modelaron por separado, y se encontraron condiciones apropiadas por medio de un método doble de superficie de respuesta. Como el estudio de caso 15.4 incluye dos modelos (media y varianza) podría considerarse un análisis doble de superficie de respuesta. Por fortuna, en este ejemplo las mismas condiciones de las

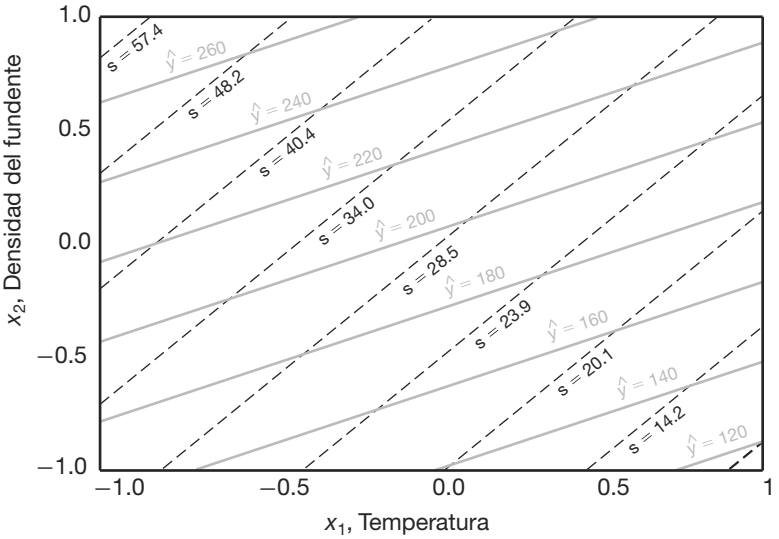


Figura 15.20: Media y desviación estándar del estudio de caso 15.4.

dos variables de control relevantes, la temperatura y la densidad del fundente, eran las óptimas para la media y la varianza del proceso. En la práctica la mayoría de las veces es necesario apelar a algún tipo de compromiso entre la media y la varianza.

El método que se ilustra en el estudio de caso 15.4 implica encontrar condiciones óptimas para el proceso cuando los datos que se utilizan provienen de un tipo de diseño experimental con arreglo de producto (o arreglo cruzado). Con frecuencia el uso de un arreglo de producto, un cruce entre dos diseños, es muy costoso. Sin embargo, el desarrollo de modelos dobles de superficie de respuesta, es decir, un modelo para la media y otro para la varianza, se puede lograr sin un arreglo de producto. El diseño que incluye tanto variables de control como de ruido se conoce como *arreglo combinado*. Este tipo de diseño y el análisis resultante se puede usar para determinar cuáles condiciones de las variables de control son las más robustas (insensibles) a la variación de las variables de ruido. Esto se puede considerar equivalente a encontrar niveles de control que minimicen la varianza del proceso producida por el movimiento de las variables de ruido.

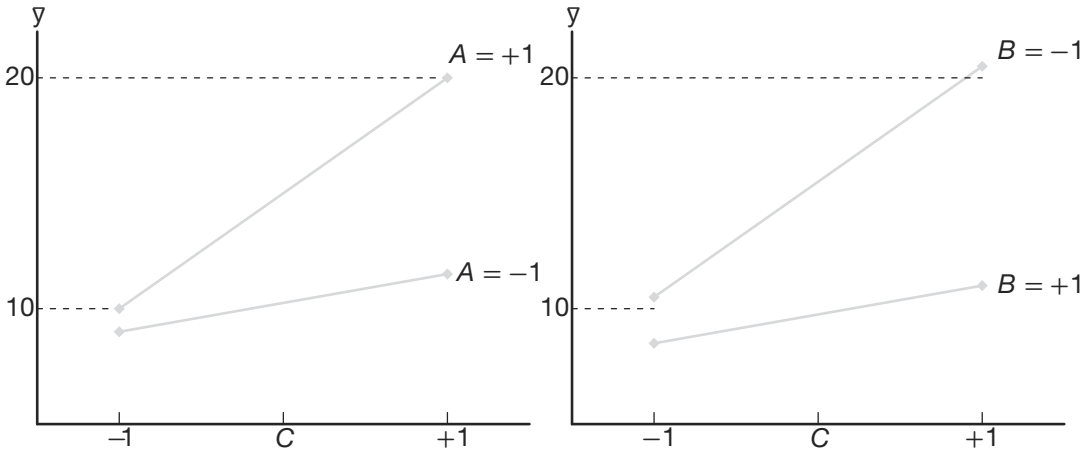
El papel de la interacción control por ruido

La estructura de la varianza del proceso es determinada en gran medida por la naturaleza de la interacción control por ruido. La naturaleza de la falta de homogeneidad de la varianza del proceso depende de cuáles variables de control interactúan con cuáles variables de ruido. De manera específica, como se ilustrará, aquellas variables de control que interactúan con una o más variables de ruido podrían ser objeto del análisis. Por ejemplo, considere un caso citado por Myers, Montgomery y Anderson-Cook (2009), el cual incluye dos variables de control y una variable de ruido con los datos que se incluyen en la tabla 15.22. *A* y *B* son las variables de control y *C* es la variable de ruido.

Tabla 15.22: Datos experimentales en un arreglo cruzado

Arreglo interno		Arreglo externo		
A	B	$C = -1$	$C = +1$	Media de respuesta
−1	−1	11	15	13.0
−1	1	7	8	7.5
1	−1	10	26	18.0
1	1	10	14	12.0

Podemos ilustrar las interacciones AC y BC con gráficas, como se observa en la figura 15.21. Es necesario entender que mientras A y B se mantienen constantes en el proceso, C sigue una distribución de probabilidad durante el mismo. Dada esta información, queda claro que $A = -1$ y $B = +1$ son niveles que producen valores más pequeños para la varianza del proceso, en tanto que $A = +1$ y $B = -1$ producen valores más grandes. Así, se dice que $A = -1$ y $B = +1$ son valores robustos, es decir, insensibles a cambios inevitables en la variable de ruido C durante el proceso.



a) Gráfica de interacción AC. *b*) Gráfica de interacción BC.

Figura 15.21: Gráficas de interacción para los datos de la tabla 15.22.

En el ejemplo anterior se dice que tanto A como B son efectos de dispersión, es decir, que ambos factores afectan la varianza del proceso. Asimismo, ambos factores son efectos de la ubicación, ya que la media de y cambia conforme los dos factores pasan de -1 a $+1$.

Análisis que incluye el modelo que contiene variables de control y de ruido

Aunque se ha hecho énfasis en que las variables de ruido no permanecen constantes durante el funcionamiento del proceso, el análisis da como resultado condiciones deseables

o incluso óptimas y proporciona de manera directa e indirecta información útil sobre el proceso. El modelo de respuesta es, en realidad, un modelo de superficie de respuesta en el vector **x** y en el vector **z**, donde **x** contiene variables de control y **z** las variables de ruido. Ciertas operaciones permiten generar modelos para la media y la varianza del proceso similares a los del estudio de caso 15.4. En Myers, Montgomery y Anderson-Cook (2009) se proporcionan los detalles; aquí se ilustrará con un ejemplo muy sencillo. Considere los datos de la tabla 15.22 de la página 650 con las variables de control A y B y la variable de ruido C. Hay ocho corridas experimentales en un factorial $2^2 \times 2$ o 2^3. Así, podemos escribir el modelo de respuesta como

$$y(x, z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z + \beta_{12} x_1 x_2 + \beta_{1z} x_1 z + \beta_{2z} x_2 z + \epsilon.$$

No se incluirán las interacciones de tres factores en el modelo de regresión. A, B y C de la tabla 15.22 están representados en el modelo por medio de x_1, x_2 y z, respectivamente. Se supone que el término del error ϵ posee las propiedades acostumbradas de independencia y varianza constante.

Las superficies de respuesta de la media y la varianza

Es más fácil comprender las superficies de respuesta de la media y la varianza del proceso si consideramos la esperanza y la varianza de z a lo largo del proceso. Se supone que la variable de ruido C [denotada por z en $y(x,z)$] es continua, con media igual a cero y varianza σ_z^2. Los modelos de la media y la varianza del proceso se pueden considerar como

$$E_z[y(x, z)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2,$$
$$\text{Var}_z[y(x, z)] = \sigma^2 + \sigma_z^2(\beta_3 + \beta_{1z} x_1 + \beta_{2z} x_2)^2 = \sigma^2 + \sigma_z^2 l_x^2,$$

donde l_x es la pendiente $\frac{\partial y(x,z)}{\partial z}$ en la dirección de z. Como se indicó antes, debemos observar que las interacciones de los factores A y B con la variable de ruido C son componentes fundamentales de la varianza del proceso.

Aunque ya se analizó este ejemplo por medio de las gráficas de la figura 15.21, las cuales revelan el papel de las interacciones AB y AC, es aleccionador ver el análisis considerando $E_z[y(x, z)]$ y $\text{Var}_z[y(x, z)]$. En este ejemplo el lector puede verificar fácilmente que el estimado b_{1z} para β_{1z} es 15/8, mientras que el estimado de b_{2z} para β_{2z} es $-15/8$. El coeficiente $b_3 = 25/8$. Así, la condición $x_1 = +1$ y $x_2 = -1$ resulta en un estimado de la varianza del proceso de

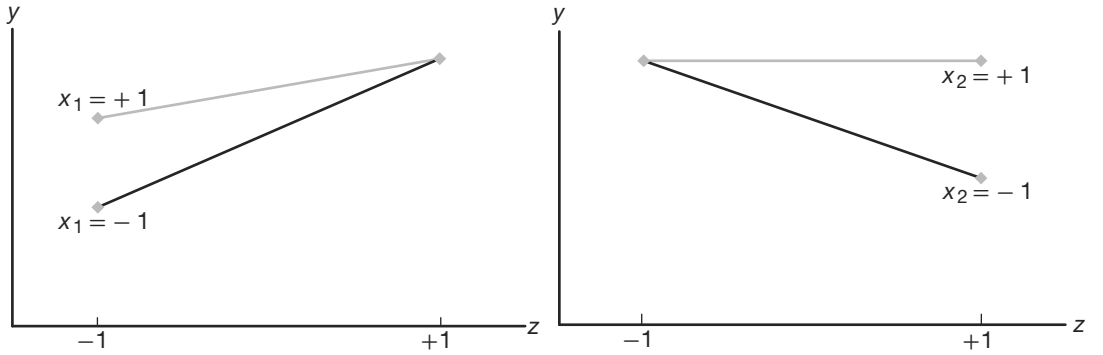
$$\widehat{\text{Var}}_z[y(x, z)] = \sigma^2 + \sigma_z^2(b_3 + b_{1z} x_1 + b_{2z} x_2)^2$$
$$= \sigma^2 + \sigma_z^2 \left[\frac{25}{8} + \left(\frac{15}{8}\right)(1) + \left(\frac{-15}{8}\right)(-1)\right]^2 = \sigma^2 + \sigma_z^2 \left(\frac{55}{8}\right)^2,$$

en tanto que para $x_1 = -1$ y $x_2 = 1$ tenemos

$$\widehat{\text{Var}}_z[y(x, z)] = \sigma^2 + \sigma_z^2(b_3 + b_{1z} x_1 + b_{2z} x_2)^2$$
$$= \sigma^2 + \sigma_z^2 \left[\frac{25}{8} + \left(\frac{15}{8}\right)(-1) + \left(\frac{15}{8}\right)(-1)\right]^2 = \sigma^2 + \sigma_z^2 \left(\frac{-5}{8}\right)^2.$$

De esta manera, para la condición más deseable (robusta) de $x_1 = -1$ y $x_2 = 1$, la varianza del proceso estimada debido a la variable de ruido C (o z) es $(25/64)\sigma_z^2$.

La condición más indeseable, la de máxima varianza del proceso, es decir, $x_1 = +1$ y $x_2 = -1$, produce una varianza del proceso estimada de $(3025/64)\sigma_z^2$. En lo que se refiere a la respuesta media, la figura 15.21 indica que si se desea una respuesta máxima, entonces $x_1 = +1$ y $x_2 = -1$ produce el mejor resultado.



a) Gráfica de interacción x_1z. *b*) Gráfica de interacción x_2z.

Figura 15.22: Gráficas de interacción para los datos del ejercicio 15.31.

Ejercicios

15.31 Considere un ejemplo en el que hay dos variables de control: x_1 y x_2, y una variable de ruido z. El objetivo consiste en determinar los niveles de x_1 y x_2, que son robustos ante los cambios de z, es decir, los niveles de x_1 y x_2 que minimizan la varianza producida en la respuesta y cuando z se mueve entre -1 a $+1$. Las variables x_1 y x_2 se encuentran a dos niveles, -1 y $+1$ en el experimento. Los datos producen las gráficas de la figura 15.22. Observe que x_1 y x_2 interactúan con la variable de ruido z. ¿Qué parámetros de x_1 y x_2 (-1 o $+1$ para cada uno) producen la varianza mínima en y? Explique sus resultados.

15.32 Considere el siguiente factorial 2^3 con variables de control x_1 y x_2 y variable de ruido z. ¿Es posible elegir x_1 y x_2 en niveles que minimicen a Var(y)? Explique su respuesta.

	$z = -1$		$z = +1$	
	$x_2 = -1$	$x_2 = +1$	$x_2 = -1$	$x_2 = +1$
$x_1 = -1$	4	6	8	10
$x_1 = +1$	1	3	3	5

15.33 Considere el estudio de caso 15.1 del moldeado por inyección. Suponga que es difícil controlar la temperatura de moldeado y, por lo tanto, que se puede asumir que en el proceso sigue una distribución normal con media igual a cero y varianza σ_z^2. El interés se centra en la varianza de la respuesta de contracción del propio proceso. Dentro del análisis de la figura 15.7 es evidente que la temperatura de moldeado, la velocidad de inyección y la interacción de ambos son los únicos factores importantes.

a) ¿El parámetro de la velocidad se podría usar para crear algún tipo de control de la varianza del proceso en la contracción que surja debido a la imposibilidad de controlar la temperatura? Explique su respuesta.

b) Utilice los estimados de parámetros de la figura 15.7 y proporcione un estimado de los siguientes modelos:
   i) contracción media a lo largo de la distribución de la temperatura;
   ii) varianza de la contracción como función de σ_z^2.

c) Utilice el modelo de la varianza estimada para determinar el nivel de velocidad que minimiza la varianza de la contracción.

d) Utilice el modelo de la contracción media para determinar qué valor de la velocidad minimiza la contracción media.

e) ¿Los resultados anteriores son consistentes con su análisis de la gráfica de interacción de la figura 15.6? Explique su respuesta.

15.34 En el estudio de caso 15.2 acerca de los datos de la purificación de carbón se sabe que el porcentaje

de sólidos en el sistema del proceso varía de manera incontrolable durante el proceso y es considerado como un factor de ruido con media igual a 0 y varianza σ_z^2. La respuesta, la eficiencia de la pureza, tiene una media y una varianza que cambian de comportamiento durante el proceso. Utilice sólo términos significativos en los siguientes incisos.

a) Utilice los estimados de la figura 15.9 para desarrollar los modelos de la varianza y la eficiencia media del proceso.
b) ¿Qué factor (o factores) podrían controlarse a ciertos niveles para controlar o minimizar la varianza del proceso?
c) ¿Qué condiciones o factores B y C dentro de la región del diseño maximizan la media estimada?
d) ¿Qué nivel de C sugeriría para minimizar la varianza del proceso cuando $B = 1$? ¿Y cuando $B = -1$?

15.35 Use los datos de purificación del carbón del ejercicio 15.2 de la página 609 para ajustar un modelo del tipo

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

donde los niveles son

x_1, porcentaje de sólidos: 8, 12
x_2, tasa de flujo: 150, 250 gal/min
x_3, pH: 5, 6

Centre y escale las variables a las unidades de diseño. Asimismo, realice una prueba para la falta de ajuste y haga comentarios acerca de lo adecuado del modelo de regresión lineal.

15.36 Se utiliza un plan factorial 2^5 para construir un modelo de regresión que contenga coeficientes de primer orden y términos del modelo para todas las interacciones de dos factores. Para cada factor se realizan corridas duplicadas. Construya la tabla de análisis de varianza que muestre los grados de libertad para la regresión, la falta de ajuste y el error puro.

15.37 Considere la fracción $\frac{1}{16}$ del factorial 2^7 que se estudió en la sección 15.9. Liste los 11 contrastes de definición adicionales.

15.38 Construya un diseño de Plackett-Burman para 10 variables que contengan 24 corridas experimentales.

Ejercicios de repaso

15.39 Se utilizó un diseño de Plackett-Burman para estudiar las propiedades reológicas de los copolímeros de alto peso molecular. En el experimento se fijaron dos niveles para cada una de seis variables. La respuesta es la viscosidad del polímero. Los datos fueron analizados en el Centro de Consultoría en Estadística de Virginia Tech, por personal del Departamento de Ingeniería Química de la universidad. Las variables son las siguientes: química del bloque duro x_1, tasa de flujo de nitrógeno x_2, tiempo de calentamiento x_3, porcentaje de compresión x_4, mediciones (alta y baja) x_5, porcentaje de esfuerzo x_6. A continuación se presentan los datos

Observación	x_1	x_2	x_3	x_4	x_5	x_6	y
1	1	−1	1	−1	−1	−1	194,700
2	1	1	−1	1	−1	−1	588,400
3	−1	1	1	−1	1	−1	7533
4	1	−1	1	1	−1	1	514,100
5	1	1	−1	1	1	−1	277,300
6	1	1	1	−1	1	1	493,500
7	−1	1	1	1	−1	1	8969
8	−1	−1	1	1	1	−1	18,340
9	−1	−1	−1	1	1	1	6793
10	1	−1	−1	−1	1	1	160,400
11	−1	1	−1	−1	−1	1	7008
12	−1	−1	−1	−1	−1	−1	3637

Construya una ecuación de regresión que relacione la viscosidad con los niveles de las seis variables. Realice pruebas t para todos los efectos principales. Recomiende los factores a conservar para estudios futuros y los factores a eliminar. Use el cuadrado medio residual (5 grados de libertad) como medida del error experimental.

15.40 Una empresa petrolera grande del suroeste lleva a cabo experimentos de manera regular para probar los aditivos de los fluidos de perforación. La viscosidad plástica es una medición reológica que refleja el espesor del fluido. Se agregan varios polímeros al fluido para incrementar su viscosidad. A continuación se presenta un conjunto de datos en el que se usaron dos polímeros, con dos niveles cada uno, y se midió la viscosidad. La concentración de los polímeros se indica como "baja" y "alta". Haga un análisis del experimento factorial 2^2. Pruebe los efectos e interacción de los dos polímeros.

	Polímero 1			
Polímero 2	**Baja**		**Alta**	
Baja	3.0	3.5	11.3	12.0
Alta	11.7	12.0	21.7	22.4

15.41 Se analiza un experimento factorial 2^2 en el Centro de Consultoría en Estadística de Virginia Tech. El cliente es miembro del Department of Housing, Interior Design, and Resource Management y le interesa comparar hornos de arranque en frío y de

precalentamiento en términos de la energía total que se entrega al producto, y, además, comparar las condiciones de convección con el modo regular. Se hicieron cuatro corridas experimentales con cada una de las cuatro combinaciones de los factores. A continuación se presentan los datos del experimento:

	Precalentamiento		Frío	
Modo de convección	618 629	619.3 611	575 574	573.7 572
Modo regular	581 581	585.7 595	558 562	562 566

Haga un análisis de varianza para estudiar la interacción y los efectos principales. Saque sus conclusiones.

15.42 En el estudio "The Use of Regression Analysis for Correcting Matrix Effects in the X-Ray Fluorescence Analysis of Pyrotechnic Compositions", publicado en *Proceedings of the Tenth Conference on the Design of Experiments in Army Research Development and Testing*, ARO-D Report 65-3 (1965), se realizó un experimento donde se hicieron variar las concentraciones de cuatro componentes de una mezcla de propulsor y los pesos de partículas finas y gruesas en la suspensión. Los factores *A*, *B*, *C* y *D*, cada uno en dos niveles, representan las concentraciones de los cuatro componentes, y los factores *E* y *F*, también en dos niveles, representan los pesos de las partículas finas y gruesas que hay en la suspensión. El objetivo del análisis era determinar si las relaciones de intensidad de rayos X asociadas con el componente 1 del propulsor eran influidas en forma significativa por la variación de las concentraciones de los distintos componentes y los pesos de las partículas, según su tamaño, en la mezcla. Se utilizó una fracción de $\frac{1}{8}$ de un experimento factorial 2^6 con los contrastes de definición *ADE*, *BCE* y *ACF*. Los datos siguientes representan el total de un par de lecturas de intensidad.

El cuadrado medio del error agrupado con 8 grados de libertad es dado por 0.02005. Analice los datos utilizando un nivel de significancia de 0.05 para determinar si las concentraciones de los componentes y los pesos de las partículas finas y gruesas presentes en la suspensión influyen de manera significativa en las relaciones de intensidad asociadas con el componente 1. Suponga que no existe interacción entre los seis factores.

Lote	Combinación de tratamientos	Relación total de intensidad
1	*abef*	2.2480
2	*cdef*	1.8570
3	(1)	2.2428
4	*ace*	2.3270
5	*bde*	1.8830
6	*abcd*	1.8078
7	*adf*	2.1424
8	*bcf*	1.9122

15.43 Utilice la tabla 15.16 para construir un diseño de 16 corridas con 8 factores que tenga resolución IV.

15.44 En el ejercicio de repaso 15.43, compruebe que el diseño en efecto tiene resolución IV.

15.45 Construya un diseño que contenga 9 puntos de diseño, sea ortogonal, contenga un total de 12 corridas y 3 grados de libertad para el error de réplica, y también que permita hacer una prueba de falta de ajuste para la curvatura cuadrática pura.

15.46 Considere un diseño 2^{3-1}_{III} con 2 corridas centrales. Considere $\bar{y}_f$ como la respuesta promedio en el parámetro de diseño y $\bar{y}_0$ como la respuesta promedio en el centro del diseño. Suponga que el verdadero modelo de la regresión es

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \\ + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2,$$

a) Proporcione (y compruebe) $E(\bar{y}_f - \bar{y}_0)$.
b) Explique lo que haya aprendido del resultado del inciso *a*

15.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

En el empleo de experimentos factoriales fraccionados uno de los aspectos más importantes que debe atender el analista es la *resolución del diseño*. Un diseño de resolución baja es más pequeño y, por lo tanto, menos costoso que uno de mayor resolución. Sin embargo, se paga un precio por el diseño más barato. El diseño de menor resolución tiene alias más pesados que uno de resolución mayor. Por ejemplo, si el investigador sospecha que las interacciones de dos factores son importantes, entonces no debería emplear la resolución III. Un diseño de resolución III es estrictamente un **plan de efectos principales**.

Capítulo 16

Estadística no paramétrica

16.1 Pruebas no paramétricas

La mayoría de los procedimientos de prueba de hipótesis que se presentaron en los capítulos anteriores se basan en la suposición de que las muestras aleatorias se seleccionan de poblaciones normales. Por fortuna la mayor parte de estas pruebas aún son confiables cuando existen ligeras desviaciones de la normalidad, en particular cuando el tamaño de la muestra es grande. Tradicionalmente, a tales procedimientos de prueba se les denomina **métodos paramétricos**. En este capítulo consideramos varios procedimientos de prueba alternativos, llamados **métodos no paramétricos** o **de distribución libre**, que a menudo no suponen conocimiento de ninguna clase acerca de las distribuciones de las poblaciones subyacentes, excepto, quizá, que éstas son continuas.

Los analistas de datos están usando procedimientos no paramétricos o de distribución libre cada vez con mayor frecuencia. En la ciencia y la ingeniería hay muchas aplicaciones en las que los datos no se reportan como valores de un continuo, sino, más bien, como una **escala ordinal** en la que es natural asignar rangos a los datos. De hecho, en este capítulo el lector notará muy pronto que los métodos de distribución libre aquí descritos implican un *análisis de rangos.* La mayoría de los analistas consideran que los cálculos involucrados en los métodos no paramétricos son muy atractivos e intuitivos.

Para revisar un ejemplo donde se aplica una prueba no paramétrica considere la situación en que dos jueces deben clasificar cinco marcas de cerveza de alta calidad asignando la categoría 1 a la marca que se considera que tiene la mejor calidad general, la categoría 2 a la segunda mejor, y así sucesivamente. Luego se puede utilizar una prueba no paramétrica para determinar si existe algún acuerdo entre los dos jueces.

También debemos señalar que las pruebas no paramétricas tienen asociadas varias desventajas. La primera es que no utilizan toda la información que proporciona la muestra, por lo tanto, cuando se pueden aplicar ambos métodos, estas últimas muestran ser menos eficientes que el procedimiento paramétrico correspondiente. En consecuencia, para lograr la misma potencia que la prueba paramétrica correspondiente, una prueba no paramétrica requerirá un tamaño muestral mayor que el que requeriría la primera.

Como antes indicamos, ligeras desviaciones de la normalidad dan como resultado desviaciones menores del ideal para las pruebas paramétricas estándar. Esto es particularmente cierto para la prueba t y la prueba F. En el caso de la prueba t y la prueba F, el

valor *P* citado podría tener un ligero error si se transgrediera moderadamente la suposición de normalidad.

En resumen, si se puede aplicar tanto una prueba paramétrica como una no paramétrica al mismo conjunto de datos, se debe aplicar la técnica paramétrica más eficiente. Sin embargo, es importante reconocer que a menudo no es posible justificar las suposiciones de normalidad, y que no siempre contamos con medidas cuantitativas. Es una ventaja que los estadísticos nos brinden diversos procedimientos no paramétricos útiles. Armado con las técnicas no paramétricas, el analista de datos tiene más herramientas para adaptar una variedad más amplia de situaciones experimentales. Se debe señalar que incluso basándose en las suposiciones de la teoría normal estándar, la eficiencia de las técnicas no paramétricas se acerca mucho más a la del procedimiento paramétrico correspondiente. Por otro lado, las grandes desviaciones de la normalidad hacen que el método no paramétrico sea mucho más eficiente que el procedimiento paramétrico.

Prueba de signo

El lector debería recordar que los procedimientos que se estudiaron en la sección 10.4 para probar la hipótesis nula de que $\mu = \mu_0$ son válidos sólo si la población es aproximadamente normal o si la muestra es grande. Sin embargo, si $n < 30$ y la población decididamente no es normal, debemos recurrir a una prueba no paramétrica.

La prueba de signo se utiliza para probar hipótesis sobre una *mediana* de la población. En el caso de muchos de los procedimientos no paramétricos, la media es reemplazada por la mediana como el **parámetro de ubicación** pertinente a probar. Recuerde que la mediana muestral se definió en la sección 1.3. El equivalente de la población, que se denota con $\tilde{\mu}$, tiene una definición análoga. Dada una variable aleatoria *X,* $\tilde{\mu}$ se define de modo que $P(X > \tilde{\mu}) \leq 0.5$ y $P(X < \tilde{\mu}) \leq 0.5$. En el caso continuo,

$$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 0.5.$$

Por supuesto, si la distribución es simétrica, la media y la mediana de la población son iguales. Al probar la hipótesis nula H_0 de que $\tilde{\mu} = \tilde{\mu}_0$ en comparación con la hipótesis alternativa adecuada, con base en una muestra aleatoria de tamaño *n,* reemplazamos cada valor de la muestra que exceda a $\tilde{\mu}_0$ con un signo *más*, y cada valor de la muestra menor que $\tilde{\mu}_0$ con un signo *menos.* Si la hipótesis nula es verdadera y la población es simétrica, la suma de los signos más debería ser casi igual a la suma de los signos menos. Cuando un signo aparece con más frecuencia de lo que debería, con base sólo en el azar, rechazamos la hipótesis de que la mediana de la población $\tilde{\mu}$ es igual a $\tilde{\mu}_0$.

En teoría, la prueba de signo sólo se puede aplicar en situaciones en las que $\tilde{\mu}_0$ no puede ser igual al valor de cualquiera de las observaciones. Aunque la probabilidad de obtener una observación muestral exactamente igual a $\tilde{\mu}_0$ cuando la población es continua es de cero, en la práctica un valor de la muestra igual a $\tilde{\mu}_0$ ocurre con frecuencia debido a una falta de precisión en el registro de los datos. Cuando se observan valores de la muestra iguales a $\tilde{\mu}_0$, se excluyen del análisis, lo cual da como resultado que se reduzca el tamaño de la muestra.

El estadístico de prueba adecuado para la prueba de signo es la variable aleatoria binomial *X*, que representa el número de signos más en la muestra aleatoria. Si la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, la probabilidad de que un valor muestral dé como resultado un signo más o uno menos es igual a $1/2$. Por lo tanto, para probar la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$, en realidad probamos la hipótesis nula de que el número de signos

más es un valor de una variable aleatoria que tiene una distribución binomial con el parámetro $p = 1/2$. Por lo tanto, los valores P para las alternativas unilateral y bilateral se pueden calcular usando esta distribución binomial. Por ejemplo, probando

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} < \tilde{\mu}_0,$$

se rechaza H_0 a favor de H_1 sólo si la proporción de signos más es lo suficientemente menor que $1/2$, es decir, cuando el valor x de la variable aleatoria es pequeño. Por lo tanto, si el valor P que se calcula

$$P = P\,(X \leq x \text{ cuando } p = 1/2)$$

es menor o igual que algún nivel de significancia α preseleccionado, se rechaza H_0 a favor de H_1. Por ejemplo, cuando $n = 15$ y $x = 3$, en la tabla A.1 encontramos que

$$P = P(X \leq 3 \text{ cuando } p = 1/2) = \sum_{x=0}^{3} b\left(x; 15, \frac{1}{2}\right) = 0.0176,$$

de manera que la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ realmente se puede rechazar a un nivel de significancia de 0.05 pero no a un nivel de 0.01.

Para probar la hipótesis

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} > \tilde{\mu}_0,$$

se rechaza H_0 a favor de H_1 sólo si la proporción de signos más es suficientemente mayor que $1/2$, es decir, cuando x es grande. En consecuencia, si el valor P calculado

$$P = P\,(X \geq x \text{ cuando } p = 1/2)$$

es menor que α, se rechaza H_0 a favor de H_1. Finalmente, para probar la hipótesis

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} \neq \tilde{\mu}_0,$$

se rechaza H_0 a favor de H_1 cuando la proporción de signos más es significativamente menor o mayor que $1/2$. Esto, por supuesto, es equivalente a que x sea tan pequeña o tan grande como se requiere. Por lo tanto, si $x < n/2$ y el valor P calculado

$$P = 2P(X \leq x \text{ cuando } p = 1/2)$$

es menor o igual que α, o si $x > n/2$ y el valor P calculado

$$P = 2P(X \geq x \text{ cuando } p = 1/2)$$

es menor o igual que α, se rechaza H_0 a favor de H_1.

Siempre que $n > 10$, las probabilidades binomiales con $p = 1/2$ se pueden aproximar a partir de la curva normal, ya que $np = nq > 5$. Suponga, por ejemplo, que deseamos probar la hipótesis

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} < \tilde{\mu}_0,$$

a un nivel de significancia $\alpha = 0.05$ para una muestra aleatoria de tamaño $n = 20$ que produce $x = 6$ signos más. Si utilizamos la aproximación de la curva normal con

$$\tilde{\mu} = np = (20)(0.5) = 10$$

y

$$\sigma = \sqrt{npq} = \sqrt{(20)(0.5)(0.5)} = 2.236,$$

encontramos que

$$z = \frac{6.5 - 10}{2.236} = -1.57.$$

Por lo tanto,

$$P = P(X \leq 6) \approx P(Z < -1.57) = 0.0582,$$

que conduce a no rechazar la hipótesis nula.

Ejemplo 16.1: Los siguientes datos representan el número de horas que funciona una desbrozadora antes de requerir una recarga:

1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.

A un nivel de significancia de 0.05 utilice la prueba de signo para probar la hipótesis de que esta desbrozadora específica funciona con una mediana de 1.8 horas antes de requerir una recarga.

***Solución*:**

1. $H_0\colon\ \tilde{\mu} = 1.8$.
2. $H_1\colon\ \tilde{\mu} \neq 1.8$.
3. $\alpha = 0.05$.
4. Estadístico de prueba: variable binomial X con $p = \frac{1}{2}$.
5. Cálculos: Al reemplazar cada valor con el símbolo "+" si excede 1.8, con el símbolo "–" si es menor que 1.8 y descartar las mediciones que sean iguales a 1.8, obtenemos la siguiente secuencia

   $$- \ + \ - \ - \ + \ - \ - \ + \ - \ -$$

   para la cual $n = 10$, $x = 3$ y $n/2 = 5$. Por lo tanto, el valor P que se obtiene de la tabla A.1 es

   $$P = 2P\left(X \leq 3 \text{ cuando } p = \frac{1}{2}\right) = 2\sum_{x=0}^{3} b\left(x; 10, \frac{1}{2}\right) = 0.3438 > 0.05.$$

6. Decisión: No se rechaza la hipótesis nula y se concluye que la mediana del tiempo de funcionamiento no difiere significativamente de 1.8 horas.

También se puede utilizar la prueba de signo para probar la hipótesis nula $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$ para observaciones de pares. Aquí se reemplaza cada diferencia, d_i, con un signo más o un signo menos, dependiendo de si la diferencia ajustada, $d_i - d_0$, es positiva o negativa. A lo largo de esta sección hemos asumido que las poblaciones son simétricas. No obstante, aun si las poblaciones fueran asimétricas, podríamos llevar a cabo el mismo procedimiento de prueba, pero las hipótesis se referirían a las medianas de la población en vez de a las medias.

Ejemplo 16.2: Una empresa de taxis intenta decidir si utilizar neumáticos radiales en vez de neumáticos regulares con cinturón le serviría para ahorrar combustible. Se equipan 16 automóviles con neumáticos radiales y se conducen por un recorrido de prueba establecido. Después se equipan los mismos automóviles con los neumáticos regulares con cinturón y se hace que los mismos conductores vuelvan a realizar el recorrido de prueba. El consumo de gasolina, en kilómetros por litro, se presenta en la tabla 16.1.

¿Podemos concluir a un nivel de significancia de 0.05 que los automóviles equipados con neumáticos radiales ahorran más combustible que los equipados con neumáticos regulares con cinturón?

Tabla 16.1: Datos para el ejemplo 16.2

Automóvil	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Neumáticos radiales	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0
Neumáticos con cinturón	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8
Automóvil	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Neumáticos radiales	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Neumáticos con cinturón	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

***Solución*:** Sean $\tilde{\mu}_1$ y $\tilde{\mu}_2$ la mediana de los kilómetros por litro para los automóviles equipados con neumáticos radiales y con cinturón, respectivamente.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 0$.

2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 > 0$.

3. $\alpha = 0.05$.

4. Estadístico de prueba: variable binomial X con $p = 1/2$.

5. Cálculos: después de reemplazar cada diferencia positiva con un símbolo "+" y cada diferencia negativa con un símbolo "–", y después de descartar las dos diferencias de cero, obtenemos la secuencia

$$+ \; - \; + \; + \; - \; + \; + \; + \; + \; + \; + \; + \; - \; +$$

para la que $n = 14$ y $x = 11$. Si usamos la aproximación de la curva normal, encontramos que

$$z = \frac{10.5 - 7}{\sqrt{(14)(0.5)(0.5)}} = 1.87,$$

y entonces

$$P = P(X \geq 11) \approx P(Z > 1.87) = 0.0307.$$

6. Decisión: Se rechaza H_0 y se concluye que, en promedio, los neumáticos radiales ahorran más combustible.

La prueba de signo no sólo es uno de los procedimientos no paramétricos más fáciles de aplicar, sino que tiene la ventaja adicional de poder aplicarse a datos dicotómicos que no se pueden registrar en una escala numérica, pero que se pueden representar mediante respuestas positivas y negativas. Por ejemplo, la prueba de signo se aplica en experimentos donde se registra una respuesta cualitativa como "éxito" o "fracaso"; y en experimentos de tipo sensorial donde se registra un signo más o un signo menos, dependiendo de si el catador del sabor identifica de manera correcta o incorrecta el ingrediente que se desea.

Intentaremos hacer comparaciones entre varios de los procedimientos no paramétricos y las pruebas paramétricas correspondientes. En el caso de la prueba de signo la competencia es, desde luego, la prueba *t*. Si se toman muestras de una distribución normal, al utilizar la prueba *t* se obtendrá como resultado la potencia más grande de la prueba. Si la distribución sólo es simétrica, aunque no sea normal, en términos de potencia se prefiere la prueba *t*, a menos que la distribución tenga "colas muy pesadas" en comparación con la distribución normal.

16.2 Prueba de rango con signo

El lector debe notar que la prueba de signo sólo utiliza los signos más y menos de las diferencias entre las observaciones y $\tilde{\mu}_0$ en el caso de una muestra, o los signos más y menos de las diferencias entre los pares de observaciones en el caso de muestras en pares; no se toma en cuenta la magnitud de esas diferencias. Una prueba que utiliza dirección y magnitud, propuesta en 1945 por Frank Wilcoxon, ahora se conoce comúnmente como **prueba de rango con signo de Wilcoxon**.

El analista puede extraer más información de los datos de manera no paramétrica si es razonable aplicar una restricción adicional a la distribución de la que se toman los datos. La prueba de rango con signo de Wilcoxon se aplica en el caso de una **distribución continua simétrica**. En esta condición se prueba la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$. Primero restamos $\tilde{\mu}_0$ de cada valor muestral y descartamos todas las diferencias iguales a cero. Las diferencias restantes se ordenan sin importar el signo. Se asigna una categoría de 1a la diferencia absoluta más pequeña, es decir, sin signo, una categoría de 2 a la siguiente más pequeña, y así sucesivamente. Cuando el valor absoluto de dos o más diferencias es el mismo, se asigna a cada uno el promedio de los rangos que se asignarían si las diferencias fueran distinguibles. Por ejemplo, si la quinta y la sexta diferencias más pequeñas tienen el mismo valor absoluto, a cada una se le asignaría una categoría de 5.5. Si la hipótesis $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, el total de los rangos que corresponden a las diferencias positivas debería ser casi igual al total de los rangos que corresponden a las diferencias negativas. Representemos estos totales con w_+ y w_-, respectivamente. Designamos el más pequeño de w_+ *y* w_- con *w.*

Al seleccionar muestras repetidas esperaríamos que w_+ y w_- y, por lo tanto, *w* variará. De esta manera, consideramos a w_+, w_- y *w* como valores de las correspondientes variables aleatorias W_+, W_- y *W*. La hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ se puede rechazar a favor de la hipótesis alternativa $\tilde{\mu} < \tilde{\mu}_0$ sólo si w_+ es pequeña y w_- es grande. De igual manera, la hipótesis alternativa $\tilde{\mu} > \tilde{\mu}_0$ se puede aceptar sólo si w_+ es grande y w_- es pequeña. Para una alternativa bilateral se puede rechazar H_0 a favor de H_1 si w_+ o w_- y, en consecuencia, *w* son suficientemente pequeñas. Por lo tanto, no importa cuál sea la hipótesis alternativa,

cuando el valor del estadístico adecuado W_+, W_- o W es suficientemente pequeño, se rechaza la hipótesis nula.

Dos muestras con observaciones en pares

Con el fin de probar la hipótesis nula de que se toman muestras de dos poblaciones simétricas continuas con $\tilde{\mu}_1 = \tilde{\mu}_2$ para el caso de muestras en pares, se ordenan las diferencias de las observaciones en pares sin importar el signo y se procede como en el caso de una sola muestra. Los diversos procedimientos de prueba para los casos de una sola muestra y de muestras en pares se resumen en la tabla 16.2.

Tabla 16.2: Prueba de rango con signo

H_0	H_1	Calcular
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	w
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	w_+
	$\tilde{\mu}_1 > \tilde{\mu}_2$	w_-
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	w

No es difícil mostrar que siempre que $n < 5$ y que el nivel de significancia no exceda a 0.05 para una prueba de una cola, o a 0.10 para una prueba de dos colas, todos los valores posibles de w_+, w_- o w conducirán a la aceptación de la hipótesis nula. Sin embargo, cuando $5 \leq n \leq 30$, la tabla A.16 muestra valores críticos aproximados de W_+ y W_- a niveles de significancia iguales a 0.01, 0.025 y 0.05 para una prueba de una cola y valores críticos de W a niveles de significancia iguales a 0.02, 0.05 y 0.10 para una prueba de dos colas. Se rechaza la hipótesis nula si el valor calculado w_+, w_- o w es **menor o igual que** el valor tabulado apropiado. Por ejemplo, cuando $n = 12$, la tabla A.16 indica que se requiere un valor de $w_+ \leq 17$ para que la hipótesis alternativa unilateral $\tilde{\mu} < \tilde{\mu}_0$ sea significativa al nivel 0.05.

Ejemplo 16.3: Repita el ejemplo 16.1 usando la prueba de rango con signo.

***Solución*:**

1. H_0: $\tilde{\mu} = 1.8$.
2. H_1: $\tilde{\mu} \neq 1.8$.
3. $\alpha = 0.05$.
4. Región crítica: Como $n = 10$, después de descartar la medida que es igual a 1.8, la tabla A.16 indica que la región crítica es $w \leq 8$.
5. Cálculos: Al restar 1.8 a cada medida y después ordenar las diferencias sin hacer caso del signo, tenemos

d_i	−0.3	0.4	−0.9	−0.5	0.2	−0.2	−0.3	0.2	−0.6	−0.1
Rangos	5.5	7	10	8	3	3	5.5	3	9	1

Ahora bien, $w_+ = 13$ y $w_- = 42$, de manera que $w = 13$, el menor de w_+ y w_-.

6. Decisión: Como antes, no se rechaza H_0 y se concluye que la mediana del tiempo de operación no difiere significativamente de 1.8 horas. ∎

La prueba de rango con signo también se puede utilizar para probar la hipótesis nula de que $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$. En este caso las poblaciones no necesitan ser simétricas. Como ocurre con la prueba de signo, restamos d_0 de cada diferencia, ordenamos las diferencias ajustadas sin importar el signo y aplicamos el mismo procedimiento anterior.

Ejemplo 16.4: Se afirma que, si se le proporcionan ejemplos de problemas con antelación, un estudiante universitario de último año puede aumentar en al menos 50 puntos su calificación en el área de especialidad del examen para ingresar a posgrado. Para probar esta afirmación se divide a un grupo de 20 estudiantes del último año en 10 pares, de manera que cada par tenga casi la misma calificación promedio durante sus 3 primeros años en la universidad. Los ejemplos de problemas y las respuestas se proporcionan al azar a un miembro de cada par una semana antes del examen. Las calificaciones del examen se presentan en la tabla 16.3.

Tabla 16.3: Datos para el ejemplo 16.4

	Par									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Con ejemplos de problemas	531	621	663	579	451	660	591	719	543	575
Sin ejemplos de problemas	509	540	688	502	424	683	568	748	530	524

A un nivel de significancia de 0.05 pruebe la hipótesis nula de que los ejemplos de problemas aumentan las calificaciones en 50 puntos, en comparación con la hipótesis alternativa de que aumentan menos de 50 puntos.

***Solución*:** Representemos con $\tilde{\mu}_1$ y $\tilde{\mu}_2$ la mediana de las calificaciones de todos los estudiantes que resuelven el examen en cuestión con y sin ejemplos de problemas, respectivamente.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 50$.

2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 < 50$.

3. $\alpha = 0.05$.

4. Región crítica: Como $n = 10$, la tabla A.16 indica que la región crítica es $w_+ \leq 11$.

5. Cálculos:

	Par									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
d_i	22	81	−25	77	27	−23	23	−29	13	51
$d_i - d_0$	−28	31	−75	27	−23	−73	−27	−79	−37	1
Rangos	5	6	9	3.5	2	8	3.5	10	7	1

Se obtiene que $w_+ = 6 + 3.5 + 1 = 10.5$.

6. Decisión: Rechazar H_0 y concluir que los ejemplos de problemas, en promedio, no aumentan las calificaciones del examen para ingresar a posgrado hasta en 50 puntos. ∎

Aproximación normal para muestras grandes

Cuando $n \geq 15$ la distribución muestral de W_+ (o W_-) se aproxima a la distribución normal, con media y varianza dadas por

$$\mu_{W_+} = \frac{n(n+1)}{4} \text{ y } \sigma^2_{W_+} = \frac{n(n+1)(2n+1).}{24}$$

Por lo tanto, cuando n excede al valor más grande en la tabla A.16 se utiliza el estadístico

$$Z = \frac{W_+ - \mu_{W_+}}{\sigma_{W_+}}$$

para determinar la región crítica para la prueba.

Ejercicios

16.1 Los siguientes datos representan el tiempo, en minutos, que un paciente tiene que esperar durante 12 visitas al consultorio de un médico antes de ser atendido:

17	15	20	20	32	28
12	26	25	25	35	24

Utilice la prueba de signo a un nivel de significancia de 0.05 para probar la afirmación del médico de que la mediana del tiempo de espera de sus pacientes no es mayor de 20 minutos.

16.2 Los siguientes datos representan el número de horas de vuelo de entrenamiento que 18 estudiantes de piloto reciben de cierto instructor antes de su primer vuelo solos:

9	12	18	14	12	14	12	10	16
11	9	11	13	11	13	15	13	14

Con las probabilidades binomiales de la tabla A.1 realice una prueba de signo a un nivel de significancia de 0.02 para probar la afirmación del instructor de que la mediana del tiempo de vuelo de entrenamiento que sus estudiantes requieren antes de volar solos es de 12 horas.

16.3 Un inspector de alimentos examina 16 latas de cierta marca de jamón para determinar el porcentaje de impurezas externas. Se registraron los siguientes datos:

2.4	2.3	3.1	2.2	2.3	1.2	1.0	2.4
1.7	1.1	4.2	1.9	1.7	3.6	1.6	2.3

Utilice una aproximación normal a la distribución binomial y realice una prueba de signo a un nivel de significancia de 0.05 para probar la hipótesis nula de que la mediana del porcentaje de impurezas en esta marca de jamón es de 2.5%, en comparación con la hipótesis alternativa de que la mediana del porcentaje de impurezas no es de 2.5%.

16.4 Un proveedor de pintura acrílica afirma que un nuevo aditivo reducirá el tiempo de secado de su pintura. Para probar esta afirmación se pintaron 12 paneles de madera; la mitad de cada panel se pintó con la pintura que contiene el aditivo regular y la otra mitad con la pintura que contiene el nuevo aditivo. Los tiempos de secado, en horas, son los siguientes:

	Tiempo de secado (horas)	
Panel	**Aditivo nuevo**	**Aditivo regular**
1	6.4	6.6
2	5.8	5.8
3	7.4	7.8
4	5.5	5.7
5	6.3	6.0
6	7.8	8.4
7	8.6	8.8
8	8.2	8.4
9	7.0	7.3
10	4.9	5.8
11	5.9	5.8
12	6.5	6.5

Utilice la prueba de signo a un nivel de 0.05 para probar la hipótesis nula de que el nuevo aditivo no disminuye el tiempo que tarda en secar la pintura con el aditivo regular.

16.5 Se afirma que una nueva dieta reducirá el peso de una persona en 4.5 kilogramos, en promedio, en un periodo de dos semanas. Se registran los pesos de 10 mujeres que siguen esta dieta, antes y después de un periodo de dos semanas, y se obtienen los siguientes datos:

Mujer	**Peso antes**	**Peso después**
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	69.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4
8	63.6	60.2
9	68.2	62.3
10	59.4	58.7

Utilice la prueba de signo a un nivel de significancia de 0.05 para probar la hipótesis de que la dieta reduce la mediana del peso en 4.5 kilogramos, en comparación con la hipótesis alternativa de que la mediana de la pérdida de peso es menor que 4.5 kilogramos.

16.6 En un experimento de contaminación atmosférica se comparan dos tipos de instrumentos para medir la cantidad de monóxido de azufre en la atmósfera. Se registraron las siguientes lecturas diarias durante dos semanas:

	Monóxido de azufre	
Día	**Instrumento *A***	**Instrumento *B***
1	0.96	0.87
2	0.82	0.74
3	0.75	0.63
4	0.61	0.55
5	0.89	0.76
6	0.64	0.70
7	0.81	0.69
8	0.68	0.57
9	0.65	0.53
10	0.84	0.88
11	0.59	0.51
12	0.94	0.79
13	0.91	0.84
14	0.77	0.63

Utilice la aproximación normal a la distribución binomial y realice una prueba de signo para determinar si los diferentes instrumentos conducen a diferentes resultados. Utilice un nivel de significancia de 0.05.

16.7 Las siguientes cifras indican la presión sanguínea sistólica de 16 corredores antes y después de una carrera de ocho kilómetros:

Corredor	**Antes**	**Después**
1	158	164
2	149	158
3	160	163
4	155	160
5	164	172
6	138	147
7	163	167
8	159	169
9	165	173
10	145	147
11	150	156
12	161	164
13	132	133
14	155	161
15	146	154
16	159	170

Utilice una prueba de signo a un nivel de significancia de 0.05 para probar la hipótesis nula de que correr ocho kilómetros aumenta la mediana de la presión sanguínea sistólica en ocho puntos, en comparación con la hipótesis alternativa de que el aumento en la mediana es menor que ocho puntos.

16.8 Analice los datos del ejercicio 16.1 usando la prueba de rango con signo.

16.9 Analice los datos del ejercicio 16.2 usando la prueba de rango con signo.

16.10 Los pesos de 5 personas, en kilogramos, antes de dejar de fumar y cinco semanas después de dejar de fumar, son los siguientes:

	Individuo				
	1	**2**	**3**	**4**	**5**
Antes	66	80	69	52	75
Después	71	82	68	56	73

Utilice la prueba de rango con signo para observaciones en pares y pruebe la hipótesis, a un nivel de significancia de 0.05, de que dejar de fumar no influye en el peso de una persona, en comparación con la hipótesis alternativa de que al dejar de fumar se aumenta de peso.

16.11 Repita el ejercicio 16.5 usando la prueba de rango con signo.

16.12 Los siguientes son los números de recetas surtidas por dos farmacias en un periodo de 20 días:

Día	**Farmacia *A***	**Farmacia *B***
1	19	17
2	21	15
3	15	12
4	17	12
5	24	16
6	12	15
7	19	11
8	14	13
9	20	14
10	18	21
11	23	19
12	21	15
13	17	11
14	12	10
15	16	20
16	15	12
17	20	13
18	18	17
19	14	16
20	22	18

A un nivel de significancia de 0.01 utilice la prueba de rango con signo para determinar si las dos farmacias surten el mismo número de recetas, “en promedio”, en comparación con la hipótesis alternativa de que la farmacia *A* surte más recetas que la farmacia *B*.

16.13 Repita el ejercicio 16.7 usando la prueba de rango con signo.

16.14 Repita el ejercicio 16.6 con la prueba de rango con signo.

16.3 Prueba de la suma de rangos de Wilcoxon

Como antes indicamos, el procedimiento no paramétrico por lo general es una alternativa adecuada para la prueba de la teoría normal cuando la suposición de normalidad no es válida. Cuando nos interesa probar la igualdad de las medias de dos distribuciones continuas que evidentemente no son normales, y las muestras son independientes, es decir, que no hay emparejamiento de observaciones, la **prueba de la suma de rangos de Wilcoxon** o la **prueba de dos muestras de Wilcoxon** es una alternativa apropiada a la prueba *t* de dos muestras que se describe en el capítulo 10.

Probaremos la hipótesis nula H_0 de que $\tilde{\mu}_1 = \tilde{\mu}_2$ en comparación con alguna hipótesis alternativa adecuada. Primero seleccionamos una muestra aleatoria de cada una de las poblaciones. Sea n_1 el número de observaciones en la muestra más pequeña y n_2 el número de observaciones en la muestra más grande. Cuando las muestras son de igual tamaño n_1 y n_2 se pueden asignar de manera aleatoria. Se ordenan las $n_1 + n_2$ observaciones de las muestras combinadas en orden ascendente y se sustituye un rango de 1, 2,..., $n_1 + n_2$ para cada observación. En el caso de empates (observaciones idénticas), se reemplazan las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. Por ejemplo, si la séptima y octava observaciones fueran idénticas, se asignaría un rango de 7.5 a cada una de las dos observaciones.

La suma de los rangos que corresponden a las n_1 observaciones en la muestra más pequeña se denota con w_1. De manera similar, el valor w_2 representa la suma de los n_2 rangos que corresponden a la muestra más grande. El total $w_1 + w_2$ depende sólo del número de observaciones en las dos muestras y de ninguna manera resulta afectado por los resultados del experimento. Por lo tanto, si $n_1 = 3$ y $n_2 = 4$, entonces $w_1 + w_2 = 1 + 2 + \cdots + 7 = 28$, sin importar los valores numéricos de las observaciones. En general,

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

la suma aritmética de los enteros 1, 2,..., $n_1 + n_2$. Una vez que se determina w_1, es más fácil calcular w_2 mediante la fórmula

$$w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - w_1.$$

Al elegir muestras repetidas de tamaños n_1 y n_2 esperaríamos que w_1 y, por lo tanto, w_2, varíen. Así, podríamos considerar a w_1 y w_2 como valores de las variables aleatorias W_1 y W_2, respectivamente. La hipótesis nula $\tilde{\mu}_1 = \tilde{\mu}_2$ se rechazará a favor de la hipótesis alternativa $\tilde{\mu}_1 < \tilde{\mu}_2$ sólo si w_1 es pequeña y w_2 es grande. De igual manera, la hipótesis alternativa $\tilde{\mu}_1 > \tilde{\mu}_2$ se puede aceptar sólo si w_1 es grande y w_2 es pequeña. Para una prueba de dos colas podemos rechazar H_0 a favor de H_1 si w_1 es pequeña y w_2 es grande, o si w_1 es grande y w_2 es pequeña. En otras palabras, se acepta la hipótesis alternativa $\tilde{\mu}_1 < \tilde{\mu}_2$ si w_1 es suficientemente pequeña; la hipótesis alternativa $\tilde{\mu}_1 > \tilde{\mu}_2$ se acepta si w_2 es suficientemente pequeña; y la hipótesis alternativa $\tilde{\mu}_1 \neq \tilde{\mu}_2$ se acepta si el mínimo de w_1 y w_2 es tan pequeño como se requiere. En la práctica real por lo general basamos nuestra decisión en el valor

$$u_1 = w_1 - \frac{n_1(n_1 + 1)}{2} \quad \text{o} \quad u_2 = w_2 - \frac{n_2(n_2 + 1)}{2}$$

del estadístico relacionado U_1 o U_2, o en el valor u del estadístico U, el mínimo de U_1 y U_2. Dichos estadísticos simplifican la construcción de tablas de valores críticos, dado

que U_1 y U_2 tienen distribuciones muestrales simétricas y toman valores en el intervalo de 0 a n_1n_2, tales que $u_1 + u_2 = n_1n_2$.

De las fórmulas para u_1 y u_2 vemos que u_1 será pequeña cuando w_1 es pequeña, y u_2 será pequeña cuando w_2 sea pequeña. En consecuencia, la hipótesis nula se rechazará siempre que los estadísticos apropiados U_1, U_2 o U tomen un valor menor o igual que el valor crítico deseado dado en la tabla A.17. Los diversos procedimientos de prueba se resumen en la tabla 16.4.

Tabla 16.4: Prueba de la suma de rangos

H_0	H_1	Calcular
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

La tabla A.17 proporciona valores críticos de U_1 y U_2 para niveles de significancia iguales a 0.001, 0.01, 0.025 y 0.05 para una prueba de una cola, y valores críticos de U para niveles de significancia iguales a 0.002, 0.02, 0.05 y 0.10 para una prueba de dos colas. Si el valor observado de u_1, u_2 o u es **menor o igual que** el valor crítico tabulado, se rechaza la hipótesis nula al nivel de significancia que se indica en la tabla. Suponga, por ejemplo, que deseamos probar la hipótesis nula de que $\tilde{\mu}_1 = \tilde{\mu}_2$ en comparación con la hipótesis alternativa unilateral de que $\tilde{\mu}_1 < \tilde{\mu}_2$ a un nivel de significancia de 0.05 para muestras aleatorias de tamaños $n_1 = 3$ y $n_2 = 5$, que producen el valor $w_1 = 8$. Se sigue que

$$u_1 = 8 - \frac{(3)(4)}{2} = 2.$$

Nuestra prueba de una sola cola se basa en el estadístico U_1. Si se usa la tabla A.17, se rechaza la hipótesis nula de medias iguales cuando $u_1 \leq 1$. Como $u_1 = 2$ no cae en la región de rechazo, no se puede rechazar la hipótesis nula.

Ejemplo 16.5: Se encontró que el contenido de nicotina de dos marcas de cigarrillos, medido en miligramos, es el siguiente:

Marca *A*	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Marca *B*	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

A un nivel de significancia de 0.05 pruebe la hipótesis de que las medianas del contenido de nicotina de las dos marcas son iguales, en comparación con la hipótesis alternativa de que son diferentes.

***Solución*:**

1. H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$.
2. H_1: $\tilde{\mu}_1 \neq \tilde{\mu}_2$.
3. $\alpha = 0.05$.
4. Región crítica: $u \leq 17$ (de la tabla A.17).
5. Cálculos: Las observaciones se acomodan en orden ascendente y se les asignan rangos del 1 al 18.

Datos originales	Rangos	Datos originales	Rangos
0.6	1	4.0	10.5*
1.6	2	4.0	10.5
1.9	3	4.1	12
2.1	4*	4.8	13*
2.2	5	5.4	14.5*
2.5	6	5.4	14.5
3.1	7	6.1	16*
3.3	8*	6.2	17
3.7	9*	6.3	18*

*Los rangos marcados con asterisco pertenecen a la muestra *A*.

Ahora

$$w_1 = 4 + 8 + 9 + 10.5 + 13 + 14.5 + 16 + 18 = 93$$

y

$$w_2 = \frac{(18)(19)}{2} - 93 = 78.$$

Por lo tanto,

$$u_1 = 93 - \frac{(8)(9)}{2} = 57, \qquad u_2 = 78 - \frac{(10)(11)}{2} = 23.$$

6. Decisión: no se rechaza la hipótesis nula H_0 y se concluye que no hay diferencia significativa en las medianas del contenido de nicotina en las dos marcas de cigarrillos. ■

Teoría normal de aproximación para dos muestras

Cuando n_1 y n_2 exceden a 8, la distribución muestral de U_1 (o U_2) se aproxima a la distribución normal con media y varianza dadas por

$$\mu_{U_1} = \frac{n_1 n_2}{2} \text{ y } \sigma^2_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

En consecuencia, cuando n_2 es mayor que 20, el valor máximo en la tabla A.17, y n_1 es al menos 9, se puede utilizar el estadístico

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

para la prueba, con la región crítica que cae ya sea en alguna o en ambas colas de la distribución normal estándar, dependiendo de la forma de H_1.

El uso de la prueba de suma de rangos de Wilcoxon no se restringe a poblaciones no normales. Se puede utilizar en vez de la prueba *t* de dos muestras cuando las poblaciones son normales, aunque la potencia será menor. La prueba de suma de rangos de Wilcoxon siempre es superior a la prueba *t* para poblaciones definitivamente no normales.

16.4 Prueba de Kruskal-Wallis

En los capítulos 13, 14 y 15 la técnica del análisis de varianza resalta como técnica analítica para probar la igualdad de $k \geq 2$ medias de la población. Sin embargo, el lector debería recordar que para que la prueba F sea teóricamente correcta se debe suponer normalidad. En esta sección investigamos una alternativa no paramétrica al análisis de varianza.

La **prueba de Kruskal-Wallis**, también llamada **prueba *H* de Kruskal-Wallis**, es una generalización de la prueba de la suma de rangos para el caso de $k > 2$ muestras. Se utiliza para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas. Presentada en 1952 por W. H. Kruskal y W. A. Wallis, la prueba constituye un procedimiento no paramétrico para probar la igualdad de las medias, en el análisis de varianza de un factor, cuando el experimentador desea evitar la suposición de que las muestras se seleccionaron de poblaciones normales.

Sea $n_i (i = 1, 2,..., k)$ el número de observaciones en la i-ésima muestra. Primero combinamos todas las k muestras y acomodamos las $n = n_1 + n_2 + \cdots + n_k$ observaciones en orden ascendente, y sustituimos el rango apropiado de 1, 2,..., n para cada observación. En el caso de empates (observaciones idénticas), seguimos el procedimiento acostumbrado de reemplazar las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. La suma de los rangos que corresponde a las n_i observaciones en la i-ésima muestra se denota mediante la variable aleatoria R_i. Consideremos ahora el estadístico

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$

que se aproxima muy bien mediante una distribución chi cuadrada con $k - 1$ grados de libertad, cuando H_0 es verdadera, siempre y cuando cada muestra conste de al menos 5 observaciones. El hecho de que h, el supuesto valor de H, sea grande cuando las muestras independientes provienen de poblaciones que no son idénticas nos permite establecer el siguiente criterio de decisión para probar H_0:

Prueba de Kruskal-Wallis

Para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas se calcula

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1),$$

donde r_i es el valor supuesto de R_i para $i = 1, 2,..., k$. Si h cae en la región crítica $H > \chi_\alpha^2$ con $v = k - 1$ grados de libertad, se rechaza H_0 al nivel de significancia α; de otra manera no se rechaza H_0.

Ejemplo 16.6: En un experimento para determinar cuál de tres diferentes sistemas de misiles es preferible, se mide la tasa de combustión del propulsor. Los datos, después de codificarlos, se presentan en la tabla 16.5. Utilice la prueba de Kruskal-Wallis y un nivel de significancia de $\alpha = 0.05$ para probar la hipótesis de que las tasas de combustión del propulsor son iguales para los tres sistemas de misiles.

Tabla 16.5: Tasas de combustión del propulsor

Sistema de misiles								
1			**2**			**3**		
24.0	16.7	22.8	23.2	19.8	18.1	18.4	19.1	17.3
19.8	18.9		17.6	20.2	17.8	17.3	19.7	18.9
						18.8	19.3	

***Solución*:**

1. H_0: $\mu_1 = \mu_2 = \mu_3$.

2. H_1: las tres medias son diferentes.

3. $\alpha = 0.05$.

4. Región crítica: $h > \chi^2_{0.05} = 5.991$, para $v = 2$ grados de libertad.

5. Cálculos: En la tabla 16.6 convertimos las 19 observaciones a rangos y sumamos los rangos para cada sistema de misiles.

Tabla 16.6: Rangos para las tasas de combustión del propulsor

Sistema de misiles		
1	**2**	**3**
19	18	7
1	14.5	11
17	6	2.5
14.5	4	2.5
9.5	16	13
$r_1 = 61.0$	5	9.5
	$r_2 = 63.5$	8
		12
		$r_3 = 65.5$

Ahora, al sustituir $n_1 = 5$, $n_2 = 6$, $n_3 = 8$ y $r_1 = 61.0$, $r_2 = 63.5$, $r_3 = 65.5$, el estadístico de prueba H toma el valor

$$h = \frac{12}{(19)(20)}\left(\frac{61.0^2}{5} + \frac{63.5^2}{6} + \frac{65.5^2}{8}\right) - (3)(20) = 1.66.$$

6. Decisión: Como $h = 1.66$ no cae en la región crítica $h > 5.991$, no hay evidencia suficiente para rechazar la hipótesis de que las tasas de combustión del propulsor son iguales para los tres sistemas de misiles.

Ejercicios

16.15 Un fabricante de cigarrillos afirma que el contenido de alquitrán de la marca de cigarrillos *B* es menor que la de la marca *A*. Para probar esta afirmación se registraron las siguientes medidas del contenido de alquitrán, en miligramos:

Marca *A*	1	12	9	13	11	14
Marca *B*	8	10	7			

Utilice la prueba de suma de rangos con $\alpha = 0.05$ para probar si la afirmación es válida.

16.16 Para averiguar si un nuevo suero detendrá la leucemia se seleccionan nueve pacientes que se encuentran en una etapa avanzada de la enfermedad. Cinco pacientes reciben el tratamiento y cuatro no. Los tiempos de supervivencia, en años, a partir del momento en que comienza el experimento son

Con tratamiento	2.1	5.3	1.4	4.6	0.9
Sin tratamiento	1.9	0.5	2.8	3.1	

Utilice la prueba de suma de rangos a un nivel de significancia de 0.05 para determinar si el suero es eficaz.

16.17 Los siguientes datos representan el número de horas que operan dos diferentes tipos de calculadoras científicas de bolsillo antes de que necesiten recargarse.

Calculadora *A*	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1
Calculadora *B*	3.8	4.8	4.3	4.2	4.0	4.9	4.5	5.2	4.5

Utilice la prueba de la suma de rangos con $\alpha = 0.01$ para determinar si la calculadora *A* opera más tiempo que la calculadora *B* con una carga completa de la batería.

16.18 Se fabrica un hilo para pesca usando dos procesos. Para determinar si hay una diferencia en la resistencia media a la rotura de los hilos, se seleccionan 10 piezas de cada proceso y después se prueba la resistencia a la rotura de cada una. Los resultados son los siguientes:

Proceso 1	10.4	9.8	11.5	10.0	9.9
	9.6	10.9	11.8	9.3	10.7
Proceso 2	8.7	11.2	9.8	10.1	10.8
	9.5	11.0	9.8	10.5	9.9

Utilice la prueba de suma de rangos con $\alpha = 0.1$ para determinar si hay diferencia entre las resistencias medias a la rotura de los hilos fabricados mediante los dos procesos.

16.19 De una clase de matemáticas de 12 estudiantes que tienen las mismas capacidades y utilizan material programado se seleccionan cinco al azar para proporcionarles enseñanza adicional. Los resultados del examen final son los siguientes:

	Calificación						
Con enseñanza adicional	87	69	78	91	80		
Sin enseñanza adicional	75	88	64	82	93	79	67

Utilice la prueba de la suma de rangos con $\alpha = 0.05$ para determinar si la enseñanza adicional influye en la calificación promedio.

16.20 Los siguientes datos representan los pesos, en kilogramos, del equipaje personal que llevan, en diferentes vuelos, un jugador de un equipo de beisbol y un jugador de un equipo de basquetbol.

Peso del equipaje (kilogramos)

Jugador de béisbol			**Jugador de basquetbol**	
16.3	20.0	18.6	15.4	16.3
18.1	15.0	15.4	17.7	18.1
15.9	18.6	15.6	18.6	16.8
14.1	14.5	18.3	12.7	14.1
17.7	19.1	17.4	15.0	13.6
16.3	13.6	14.8	15.9	16.3
13.2	17.2	16.5		

Utilice la prueba de la suma de rangos con $\alpha = 0.05$ para probar la hipótesis nula de que los dos atletas llevan la misma cantidad de equipaje en promedio, en comparación con la hipótesis alternativa de que el peso promedio del equipaje de los dos atletas es diferente.

16.21 Los siguientes datos representan los tiempos de funcionamiento, en horas, para tres tipos de calculadoras científicas de bolsillo, antes de que requieran recarga:

Calculadora

A			***B***			***C***		
4.9	6.1	4.3	5.5	5.4	6.2	6.4	6.8	5.6
4.6	5.2		5.8	5.5	5.2	6.5	6.3	6.6
				4.8				

Utilice la prueba de Kruskal-Wallis a un nivel de significancia de 0.01, para probar la hipótesis de que los tiempos de funcionamiento de las tres calculadoras son iguales.

16.22 En el ejercicio 13.6 de la página 519 utilice la prueba de Kruskal-Wallis, a un nivel de significancia de 0.05, para determinar si los solventes químicos orgánicos difieren de manera significativa en su tasa de absorción.

16.5 Pruebas de rachas

Al aplicar los diversos conceptos estadísticos que se presentan a lo largo de este libro siempre asumimos que los datos muestrales se reunieron mediante algún procedimiento aleatorio. Las **pruebas de rachas**, que se basan en el orden en el que se obtienen las observaciones muestrales, constituyen una técnica útil para probar la hipótesis nula H_0 de que las observaciones en realidad se obtuvieron al azar.

Para ilustrar las pruebas de rachas suponga que se encuesta a 12 personas para saber si utilizan cierto producto. Se cuestionaría seriamente la supuesta aleatoriedad de la muestra si las 12 personas fueran del mismo sexo. Designaremos a un hombre y a una mujer con los símbolos H y M, respectivamente, y registraremos los resultados de acuerdo con su género en el orden en que ocurren. Una secuencia común para el experimento sería

$$\underbrace{M\ M}\ \underbrace{F\ F\ F}\ \underbrace{M}\ \underbrace{F\ F}\ \underbrace{M\ M\ M\ M},$$

donde agrupamos las subsecuencias de símbolos idénticos. Tales agrupamientos se llaman **rachas**.

Definición 16.1: Una **racha** es una subsecuencia de uno o más símbolos idénticos que representan una propiedad común de los datos.

Sin importar si las mediciones de la muestra representan datos cualitativos o cuantitativos, la prueba de rachas divide los datos en dos categorías mutuamente excluyentes: hombre o mujer, defectuoso o no defectuoso, cara o cruz, arriba o abajo de la mediana, etcétera. En consecuencia, una secuencia siempre estará limitada a dos símbolos distintos. Sea n_1 el número de símbolos asociados con la categoría de menor ocurrencia, y n_2 el número de símbolos que pertenecen a la otra categoría. Entonces, el tamaño de la muestra $n = n_1 + n_2$.

Para los $n = 12$ símbolos en nuestra encuesta tenemos cinco rachas, donde la primera incluye dos H, la segunda tres M, y así sucesivamente. Si el número de rachas es mayor o menor que el que esperaríamos por el azar, se debe rechazar la hipótesis de que la muestra se extrajo al azar. Ciertamente, una muestra que tiene como resultado sólo dos corridas,

$$H\ H\ H\ H\ H\ H\ H\ M\ M\ M\ M\ M$$

o la inversa, es muy improbable que provenga de un proceso de selección aleatorio. Este resultado indicaría que las primeras siete personas entrevistadas son hombres, seguidos de cinco mujeres. Asimismo, si la muestra tiene como resultado el número máximo de 12 rachas, como en la secuencia alternada

$$H\ M\ H\ M\ H\ M\ H\ M\ H\ M\ H\ M,$$

de nuevo sospecharíamos del orden en que se seleccionaron los individuos para la encuesta.

La prueba de rachas para la aleatoriedad se basa en la variable aleatoria V, el número total de rachas que suceden en la secuencia completa del experimento. En la tabla A.18 se dan valores de $P(V \leq v^*$ cuando H_0 es verdadera) para $v^* = 2, 3, ..., 20$ rachas y valores

de n_1 y n_2 menores o iguales que 10. Los valores P tanto para pruebas de una cola como de dos colas se pueden obtener usando estos valores tabulados.

En la encuesta anterior presentamos un total de 5 *M* y 7 *H*. De aquí, con $n_1 = 5$, $n_2 = 7$ y $v = 5$, en la tabla A.18 observamos que el valor P para una prueba de dos colas es

$$P = 2P(V \leq 5 \text{ cuando } H_0 \text{ es verdadera}) = 0.394 > 0.05.$$

Es decir, el valor $v = 5$ es razonable a un nivel de significancia de 0.05 cuando H_0 es verdadera y, por lo tanto, no tenemos suficiente evidencia para rechazar la hipótesis de aleatoriedad de nuestra muestra.

Cuando el número de rachas es grande, por ejemplo, cuando $v = 11$ y $n_1 = 5$ y $n_2 = 7$, entonces el valor P en una prueba de dos colas es

$$\begin{aligned} P &= 2P(V \geq 11 \text{ cuando } H_0 \text{ es verdadera}) = 2[1 - P(V \leq 10 \text{ cuando } H_0 \text{ es verdadera})] \\ &= 2(1 - 0.992) = 0.016 < 0.05, \end{aligned}$$

que nos lleva a rechazar la hipótesis de que los valores de la muestra ocurren al azar.

La prueba de rachas también sirve para detectar desviaciones en la aleatoriedad de una secuencia de mediciones cuantitativas a lo largo del tiempo, ocasionadas por tendencias o periodos. Al reemplazar cada medición en el orden en que se obtiene, con un símbolo *más* si caen por arriba de la mediana, o con un símbolo *menos* si caen por debajo de la mediana, y omitiendo todas las mediciones que son exactamente iguales a la mediana, se genera una secuencia de signos de más y menos que se somete a prueba para verificar su aleatoriedad, como se ilustra en el siguiente ejemplo.

Ejemplo 16.7: Una máquina vierte adelgazador de pintura acrílica en un contenedor. ¿Si se mide el contenido de los siguientes 15 contenedores y los resultados son 3.6, 3.9, 4.1, 3.6, 3.8, 3.7, 3.4, 4.0, 3.8, 4.1, 3.9, 4.0, 3.8, 4.2 y 4.1 litros, diría que la cantidad de adelgazador de pintura que despacha la máquina varía de forma aleatoria? Utilice un nivel de significancia de 0.1.

***Solución*:**

1. H_0: La secuencia es aleatoria.
2. H_1: La secuencia no es aleatoria.
3. $\alpha = 0.1$.
4. Estadístico de prueba: V, número total de rachas.
5. Cálculos: Para la muestra dada encontramos $\bar{x} = 3.9$. Al reemplazar cada medición por el símbolo "+", si cae por arriba de 3.9, por el signo "–" si cae por debajo de 3.9, y si se omiten las dos mediciones que son iguales a 3.9, obtenemos la secuencia

   $$- \; + \; - \; - \; - \; - \; + \; - \; + \; + \; - \; + \; +$$

   para la que $n_1 = 6$, $n_2 = 7$ y $v = 8$. Por lo tanto, de la tabla A.18, el valor P calculado es

   $$\begin{aligned} P &= 2P(V \geq 8 \text{ cuando } H_0 \text{ es verdadera}) \\ &= 2[1 - P(V \leq 8 \text{ cuando } H_0 \text{ es verdadera})] = 2(0.5) = 1. \end{aligned}$$

6. Decisión: No se rechaza la hipótesis de que la secuencia de mediciones varía de forma aleatoria.

La prueba de rachas, aunque menos poderosa, también se utiliza como una alternativa a la prueba de dos muestras de Wilcoxon para probar la afirmación de que dos muestras aleatorias provienen de poblaciones que tienen la misma distribución y, por lo tanto, medias iguales. Si las poblaciones son simétricas, el rechazo de la afirmación de distribuciones iguales es equivalente a aceptar la hipótesis alternativa de que las medias no son iguales. Para hacer la prueba primero se combinan las observaciones de ambas muestras y se acomodan en orden ascendente. Ahora se asigna la letra *A* a cada observación tomada de una de las poblaciones, y la letra *B* a cada observación de la otra población, generando así una secuencia que consta de los símbolos *A* y *B*. Si las observaciones de una población se vinculan con las observaciones de la otra población, la secuencia de símbolos *A* y *B* que se genera no será única y, en consecuencia, es poco probable que el número de rachas sea único. Los procedimientos para romper los empates por lo general dan como resultado tediosos cálculos adicionales, por lo que siempre que ocurran dichas situaciones sería preferible aplicar la prueba de la suma de rangos de Wilcoxon.

Con el fin de ilustrar el uso de las rachas al probar la igualdad de medias, considere los tiempos de supervivencia de los pacientes de leucemia del ejercicio 16.16 de la página 670, para los que tenemos

0.5	0.9	1.4	1.9	2.1	2.8	3.1	4.6	5.3
B	*A*	*A*	*B*	*A*	*B*	*B*	*A*	*A*

que resultan en $v = 6$ rachas. Si las dos poblaciones simétricas tienen medias iguales, las observaciones de las dos muestras estarán entremezcladas, lo cual dará como resultado muchas rachas. Sin embargo, si las medias de la población son significativamente diferentes, esperaríamos que la mayoría de las observaciones de una de las dos muestras fueran más pequeñas que las de la otra muestra. En el caso extremo de que las poblaciones no se traslapen, obtendríamos una secuencia de la forma

$$A\ A\ A\ A\ A\ B\ B\ B\ B \quad \text{o} \quad B\ B\ B\ B\ A\ A\ A\ A\ A$$

y en cualquier caso sólo habría dos rachas. En consecuencia, la hipótesis de medias de la población iguales se rechazará a un nivel de significancia α sólo cuando v sea suficientemente pequeña, de modo que

$$P = P(V \leq v \text{ cuando } H_0 \text{ es verdadera}) \leq \alpha,$$

lo que implica una prueba de una cola.

Si regresamos a los datos del ejercicio 16.16 de la página 670, para los que $n_1 = 4$, $n_2 = 5$ y $v = 6$, en la tabla A.18 encontramos que

$$P = P(V \leq 6 \text{ cuando } H_0 \text{ es verdadera}) = 0.786 > 0.05$$

y, por lo tanto, no se rechaza la hipótesis nula de medias iguales. De aquí concluimos que el nuevo suero no prolonga la vida, ya que no detiene la leucemia.

Cuando n_1 y n_2 aumentan en tamaño, la distribución de muestreo de V se aproxima a la distribución normal con media y varianza dadas por

$$\mu_V = \frac{2n_1n_2}{n_1 + n_2} + 1 \text{ y } \sigma_V^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

En consecuencia, cuando n_1 y n_2 son ambos mayores que 10, se puede utilizar el estadístico

$$Z = \frac{V - \mu_V}{\sigma_V}$$

con el fin de establecer la región crítica para la prueba de rachas.

16.6 Límites de tolerancia

En el capítulo 9 se analizaron los límites de tolerancia para una distribución normal de mediciones. En esta sección consideramos un método para construir intervalos de tolerancia que sean independientes de la forma de la distribución subyacente. Como se podría sospechar, para un grado de confianza razonable serán considerablemente más grandes que los que se construyen cuando se supone normalidad, y el tamaño de la muestra que se requiere es por lo general muy grande. Los límites de tolerancia no paramétricos se establecen en términos de las observaciones más grande y más pequeña en nuestra muestra.

Límites de tolerancia bilaterales

Para cualquier distribución de mediciones los límites de tolerancia bilaterales son indicados por las observaciones más grande y más pequeña en una muestra de tamaño n, donde n se determina de manera que se asegure, con $100(1-\gamma)\%$ de confianza, que **al menos** la proporción $1-\alpha$ de la distribución está incluida entre los extremos de la muestra.

La tabla A.19 proporciona los tamaños de la muestra requeridos para los valores seleccionados de γ y $1-\alpha$. Por ejemplo, cuando $\gamma = 0.01$ y $1-\alpha = 0.95$, debemos seleccionar una muestra aleatoria de tamaño $n = 130$ para tener 99% de confianza en que al menos 95% de la distribución de mediciones está incluido entre los extremos de la muestra.

En vez de determinar un tamaño muestral n tal que una proporción específica de mediciones esté contenida entre los extremos de la muestra, en muchos procesos industriales es deseable determinar un tamaño de la muestra tal que una proporción fija de la población caiga por debajo de la observación más grande (o por arriba de la más pequeña) de la muestra. Tales límites se denominan límites de tolerancia unilaterales.

Límites de tolerancia unilaterales

Para cualquier distribución de mediciones un límite de tolerancia unilateral se determina mediante la observación más pequeña (o más grande) en una muestra de tamaño n, donde n se determina de manera que se pueda asegurar con $100(1-\gamma)\%$ de confianza que **al menos** la proporción $1-\alpha$ de la distribución excederá a la observación más pequeña (menor que la mayor) de la muestra.

La tabla A.20 muestra los tamaños de la muestra requeridos, correspondientes a valores seleccionados de γ y $1-\alpha$. De aquí, cuando $\gamma = 0.05$ y $1-\alpha = 0.70$, debemos elegir una muestra de tamaño $n = 9$ para tener 95% de confianza en que 70% de nuestra distribución de mediciones excederá la observación más pequeña de la muestra.

16.7 Coeficiente de correlación de rango

En el capítulo 11 utilizamos el coeficiente de correlación muestral r para medir el coeficiente de correlación poblacional ρ, la relación lineal entre dos variables continuas X y Y. Si los rangos 1, 2,..., n se asignan a las observaciones x en orden de magnitud y de manera similar a las observaciones y, y si estos rangos se sustituyen después con los valores numéricos reales en la fórmula para el coeficiente de correlación del capítulo 11, obtenemos el equivalente no paramétrico del coeficiente de correlación convencional. Un coeficiente de correlación calculado de esta forma se conoce como **coeficiente de correlación de rangos de Spearman** y se denota con r_s. Cuando no hay empates entre ambos conjuntos de mediciones la fórmula para r_s se reduce a una expresión mucho más simple que incluye las diferencias d_i entre los rangos asignados a los n pares de x y y que establecemos ahora.

Coeficiente de correlación de rangos

Una medida no paramétrica de la asociación entre dos variables X y Y es dada por el **coeficiente de correlación de rango**

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2,$$

donde d_i es la diferencia entre los rangos asignados a x_i y y_i, y n es el número de pares de datos.

En la práctica, la fórmula anterior también se usa cuando hay empates entre las observaciones x o y. Los rangos para observaciones empatadas se asignan de la misma manera que en la prueba de rango con signo al promediar los rangos que se habrían asignado si las observaciones fueran distinguibles.

El valor de r_s por lo general se acercará al valor que se obtiene al calcular r con base en mediciones numéricas y se interpreta de forma muy similar. Como antes, el valor de r_s irá de -1 a $+1$. Un valor de $+1$ o -1 indica una asociación perfecta entre X y Y; el signo más ocurre para rangos idénticos y el signo menos para rangos inversos. Cuando r_s se acerca a cero, se concluye que las variables no están correlacionadas.

Ejemplo 16.8: Las cifras que se listan en la tabla 16.7, publicadas por la Comisión Federal de Comercio, muestran los miligramos de alquitrán y nicotina que se encontraron en 10 marcas de cigarrillos. Calcule el coeficiente de correlación de rangos para medir el grado de relación entre el contenido de alquitrán y de nicotina en cigarrillos.

Tabla 16.7: Contenidos de alquitrán y nicotina

Marca de cigarrillos	Contenido de alquitrán	Contenido de nicotina
Viceroy	14	0.9
Marlboro	17	1.1
Chesterfield	28	1.6
Kool	17	1.3
Kent	16	1.0
Raleigh	13	0.8
Old Gold	24	1.5
Philip Morris	25	1.4
Oasis	18	1.2
Players	31	2.0

***Solución*:** Sean X y Y los contenidos de alquitrán y nicotina, respectivamente. Primero asignamos rangos a cada conjunto de medidas, con el rango de 1 asignado al número más bajo en cada conjunto, el rango de 2 al segundo número más bajo en cada conjunto, y así sucesivamente, hasta que se asigna el rango 10 al número más grande. La tabla 16.8 muestra los rangos individuales de las mediciones y las diferencias en rangos para los 10 pares de observaciones.

Tabla 16.8: Rangos para los contenidos de alquitrán y nicotina

Marca de cigarrillos	x_i	y_i	d_i
Viceroy	2.0	2.0	0.0
Marlboro	4.5	4.0	0.5
Chesterfield	9.0	9.0	0.0
Kool	4.5	6.0	−1.5
Kent	3.0	3.0	0.0
Raleigh	1.0	1.0	0.0
Old Gold	7.0	8.0	−1.0
Philip Morris	8.0	7.0	1.0
Oasis	6.0	5.0	1.0
Players	10.0	10.0	0.0

Al sustituir en la fórmula para r_s, encontramos que

$$r_s = 1 - \frac{(6)(5.50)}{(10)(100-1)} = 0.967,$$

lo que indica una correlación positiva alta entre las cantidades de alquitrán y de nicotina que se encuentra en los cigarrillos. ■

Hay algunas ventajas al usar r_s en vez de r. Por ejemplo, ya no suponemos que la relación fundamental entre X y Y es lineal, por lo tanto, cuando los datos poseen una relación curvilínea distinta, el coeficiente de correlación de rangos probablemente será más confiable que la medida convencional. Una segunda ventaja del uso del coeficiente de correlación de rangos es el hecho de que no se hacen suposiciones de normalidad respecto a las distribuciones de X y Y. Quizá la mayor ventaja ocurre cuando no somos capaces de hacer mediciones numéricas significativas y, sin embargo, se pueden establecer rangos. Tal es el caso, por ejemplo, cuando diferentes jueces clasifican a un grupo de individuos de acuerdo con algún atributo. El coeficiente de correlación de rangos se puede utilizar en esta situación como una medida de la consistencia de los dos jueces.

Para probar la hipótesis de que $\rho = 0$ utilizando un coeficiente de correlación de rangos, se necesita considerar la distribución muestral de los valores r_s, con base en la suposición de que no hay correlación. En la tabla A.21 aparecen valores críticos calculados para $\alpha = 0.05, 0.025, 0.01$ y 0.005. La elaboración de esta tabla es similar a la tabla de valores críticos para la distribución t, excepto por la columna izquierda, que ahora proporciona el número de pares de observaciones en vez de los grados de libertad. Como la distribución de los valores r_s es simétrica alrededor de cero cuando $\rho = 0$, el valor r_s que deja una área de α a la izquierda es igual al negativo del valor r_s que deja una área de α a la derecha. Para una hipótesis alternativa bilateral la región crítica de tamaño α cae igualmente en las dos colas de la distribución. Para una prueba en la que la hipótesis alternativa es negativa, la región crítica está completamente en la cola izquierda de la distribución y, cuando la hipótesis alternativa es positiva, la región crítica se coloca por completo en la cola derecha.

Ejemplo 16.9: Remítase al ejemplo 16.8 y pruebe la hipótesis de que la correlación entre la cantidad de alquitrán y nicotina encontrada en los cigarrillos es cero en comparación con la hipótesis alternativa de que es mayor que cero. Utilice un nivel de significancia de 0.01.

***Solución*:**

1. H_0: $\rho = 0$.
2. H_1: $\rho > 0$.
3. $\alpha = 0.01$.
4. Región crítica: $r_s > 0.745$, de la tabla A.21.
5. Cálculos: Del ejemplo 16.8, $r_s = 0.967$.
6. Decisión: Se rechaza H_0 y se concluye que hay una correlación significativa entre la cantidad de alquitrán y nicotina que se encuentra en los cigarrillos. ∎

Con base en la suposición de que no hay correlación, se puede demostrar que la distribución de los valores rs se aproxima a una distribución normal, con una media igual a cero y una desviación estándar de $1/\sqrt{n-1}$ conforme aumenta n. En consecuencia, cuando n excede a los valores dados en la tabla A.21 se puede probar si existe una correlación significativa calculando

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$

y comparando con los valores críticos de la distribución normal estándar que se presentan en la tabla A.3.

Ejercicios

16.23 Con el fin de estimar la proporción de votantes que favorecen a cierto candidato para alcalde, se selecciona una muestra aleatoria de 15 adultos que viven en una pequeña ciudad. También se le pregunta a cada individuo si se graduó de la universidad. Al denotar con *S* y *N* las respuestas "sí" y "no", respectivamente, a la pregunta sobre la escolaridad, se obtuvo la siguiente secuencia:

N N N N Y Y N Y Y N Y N N N N

Utilice la prueba de rachas a un nivel de significancia de 0.1 para determinar si la secuencia apoya la afirmación de que la muestra se seleccionó al azar.

16.24 Se utiliza un proceso de plateado para cubrir cierto tipo de charola de servicio. Cuando el proceso está bajo control el espesor de la plata sobre la charola variará de forma aleatoria siguiendo una distribución normal con una media de 0.02 milímetros y una desviación estándar de 0.005 milímetros. Suponga que las siguientes 12 charolas examinadas muestran los siguientes espesores de plata: 0.019, 0.021, 0.020, 0.019, 0.020, 0.018, 0.023, 0.021, 0.024, 0.022, 0.023, 0.022. Utilice la prueba de rachas para determinar si las fluctuaciones en el espesor de una charola a otra son aleatorias. Utilice $\alpha = 0.05$.

16.25 Use la prueba de rachas a un nivel de 0.01 para probar si hay una diferencia en el tiempo promedio de funcionamiento para las dos calculadoras del ejercicio 16.17 de la página 670.

16.26 En una línea de producción industrial los artículos se inspeccionan de forma periódica en busca de defectos. La siguiente es una secuencia de artículos defectuosos, *D,* y no defectuosos, *N,* producidos por esta línea:

D D N N N D N N D D N N N N
N D D D N N D N N N N D N D

Utilice la teoría de muestras grandes para la prueba de rachas a un nivel de significancia de 0.05 para determinar si los artículos defectuosos ocurren al azar.

16.27 Suponga que las mediciones del ejercicio 1.14 de la página 30 se registraron en renglones sucesivos de izquierda a derecha conforme se reunieron. Utilice la prueba de rachas con $\alpha = 0.05$ para probar la hipótesis de que los datos representan una secuencia aleatoria.

16.28 ¿Qué tan grande debe ser una muestra para tener 95% de confianza en que al menos 85% de la distribución de medidas se incluye entre los extremos de la muestra?

16.29 ¿Cuál es la probabilidad de que el rango de una muestra aleatoria de tamaño 24 incluya al menos a 90% de la población?

16.30 ¿Qué tan grande debe ser una muestra para tener 99% de confianza en que al menos 80% de la población será menor que la observación más grande de la muestra?

16.31 ¿Cuál es la probabilidad de que al menos 95% de una población exceda al valor más pequeño en una muestra aleatoria de tamaño $n = 135$?

16.32 En la siguiente tabla se presentan las calificaciones registradas de 10 estudiantes en un examen de medio curso y las del examen final en un curso de cálculo:

Estudiante	Examen de medio curso	Examen final
L.S.A.	84	73
W.P.B.	98	63
R.W.K.	91	87
J.R.L.	72	66
J.K.L.	86	78
D.L.P.	93	78
B.L.P.	80	91
D.W.M.	0	0
M.N.M.	92	88
R.H.S.	87	77

a) Calcule el coeficiente de correlación de rangos.
b) Pruebe la hipótesis nula de que $\rho = 0$ en comparación con la hipótesis alternativa de que $\rho > 0$. Utilice $\alpha = 0.025$.

16.33 Refiérase a los datos del ejercicio 11.1 de la página 398 y
a) calcule el coeficiente de correlación de rangos;
b) a un nivel de significancia de 0.05 pruebe la hipótesis nula de que $\rho = 0$, en comparación con la hipótesis alternativa de que $\rho \neq 0$. Compare sus resultados con los obtenidos en el ejercicio 11.44 de la página 435.

16.34 Calcule el coeficiente de correlación de rangos para la precipitación pluvial diaria y la cantidad de partículas eliminadas en el ejercicio 11.13 de la página 400.

16.35 Refiérase a los datos del ejercicio 11.47 de la página 436 respecto al peso y tamaño de tórax de los bebés, y
a) calcule el coeficiente de correlación de rangos;
b) a un nivel de significancia de 0.025, pruebe la hipótesis de que $\rho = 0$ en comparación con la hipótesis alternativa de que $\rho > 0$.

16.36 Un grupo de consumidores prueba la calidad general de nueve marcas de hornos de microondas. Los rangos asignados por el grupo y los precios de venta al menudeo sugeridos son los siguientes:

Fabricante	Clasificación del grupo	Precio sugerido ($)
A	6	480
B	9	395
C	2	575
D	8	550
E	5	510
F	1	545
G	7	400
H	4	465
I	3	420

¿Existe una relación significativa entre la calidad y el precio de un horno de microondas? Utilice un nivel de significancia de 0.05.

16.37 En un desfile de regreso a clases dos jueces califican ocho carros alegóricos en el siguiente orden:

	Carro alegórico							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Juez *A*	5	8	4	3	6	2	7	1
Juez *B*	7	5	4	2	8	1	6	3

a) Calcule el coeficiente de correlación de rangos.
b) Pruebe la hipótesis nula de que $\rho = 0$ en comparación con la hipótesis alternativa de que $\rho > 0$. Use $\alpha = 0.05$.

16.38 En el artículo titulado "Risky Assumptions" de Paul Slovic, Baruch Fischoff y Sarah Lichtenstein, publicado en *Psychology Today* (junio de 1980), miembros de la Liga de Mujeres Votantes y expertos profesionalmente implicados en la evaluación de riesgos clasificaron el riesgo de muerte, en Estados Unidos, de realizar 30 actividades y utilizar tecnologías. Las puntuaciones se presentan en la tabla 16.9.
a) Calcule el coeficiente de correlación de rangos.
b) Pruebe la hipótesis nula de cero correlación entre las clasificaciones de la Liga de Mujeres Votantes y de los expertos en comparación con la hipótesis alternativa de que la correlación no es igual a cero. Utilice un nivel de significancia de 0.05.

Tabla 16.9: Rango de datos para el ejercicio 16.38

Riesgo de la actividad o tecnología	Votantes	Expertos	Riesgo de la actividad o tecnología	Votantes	Expertos
Energía nuclear	1	20	Vehículos de motor	2	1
Armas de fuego	3	4	Tabaquismo	4	2
Motocicletas	5	6	Bebidas alcohólicas	6	3
Aviación privada	7	12	Trabajo policiaco	8	17
Pesticidas	9	8	Cirugía	10	5
Bombero	11	18	Construcción grande	12	13
Cacería	13	23	Latas de aerosol	14	26
Montañismo	15	29	Bicicletas	16	15
Aviación comercial	17	16	Energía eléctrica	18	9
Natación	19	10	Anticonceptivos	20	11
Esquí	21	30	Rayos X	22	7
Futbol americano	23	27	Ferrocarriles	24	19
Conservadores de alimentos	25	14	Colorantes de alimentos	26	21
Podadoras	27	28	Antibióticos	28	24
Electrodomésticos	29	22	Vacunas	30	25

Ejercicios de repaso

16.39 Un estudio de una empresa química compara las propiedades de desecación de dos diferentes polímeros. Se utilizaron 10 lodos diferentes y se permitió que ambos polímeros secaran cada lodo. El secado libre se midió en mL/min.

Tipo de lodo	Polímero *A*	Polímero *B*
1	12.7	12.0
2	14.6	15.0
3	18.6	19.2
4	17.5	17.3
5	11.8	12.2
6	16.9	16.6
7	19.9	20.1
8	17.6	17.6
9	15.6	16.0
10	16.0	16.1

a) Utilice la prueba de signos a un nivel de 0.05 para probar la hipótesis nula de que el polímero *A* tiene la misma mediana de secado que el polímero *B*.
b) Utilice la prueba de rangos con signo para probar la hipótesis del inciso *a*.

16.40 En el ejercicio de repaso 13.45 de la página 555 use la prueba de Kruskal-Wallis, a un nivel de significancia de 0.05, para determinar si los análisis químicos realizados por los cuatro laboratorios producen, en promedio, los mismos resultados.

16.41 Use los datos del ejercicio 13.14 de la página 530 para ver si la cantidad mediana de pérdida de nitrógeno en la transpiración difiere para los tres niveles de proteína dietética.

Capítulo 17

Control estadístico de la calidad

17.1 Introducción

La idea de usar técnicas de muestreo y análisis estadístico en un entorno de producción tuvo sus comienzos en la década de 1920. El objetivo de este concepto tan exitoso es reducir de manera sistemática la variabilidad y el aislamiento asociados con las fuentes de dificultades *durante la producción.* En 1924 Walter A. Shewhart, de la empresa Bell Telephone Laboratories, desarrolló el concepto de gráfica de control. Sin embargo, fue hasta la Segunda Guerra Mundial cuando se generalizó el uso de este tipo de gráficas debido a la importancia que durante ese periodo tuvo el mantenimiento de la calidad en los procesos de producción. En las décadas de 1950 y 1960 el desarrollo del control de calidad y el área general de seguridad de la calidad crecieron con rapidez, en particular con el surgimiento del programa espacial en Estados Unidos. En Japón hubo un amplio y exitoso uso del control de calidad gracias a los esfuerzos de W. Edwards Deming, quien trabajó como consultor en Japón después de la Segunda Guerra Mundial. El control de calidad ha sido, y es, un elemento importante en el desarrollo de la industria y la economía de Japón.

El control de calidad está recibiendo cada vez más atención como una herramienta de administración en la cual se observan y evalúan las características importantes de un producto en comparación con algún tipo de estándar. Los diversos procedimientos en el control de calidad implican un uso considerable de los procedimientos de muestreo y los principios estadísticos expuestos en capítulos anteriores. Los principales usuarios del control de calidad son, por supuesto, las corporaciones industriales. Es evidente que un programa eficaz de control de calidad mejora la calidad del artículo que se produce y aumenta las utilidades. Esto es particularmente cierto en la actualidad, pues los productos se fabrican en volúmenes altos. Antes de que surgiera el movimiento hacia los métodos de control de calidad, a menudo ésta se veía afectada debido a la falta de eficiencia, lo cual, por supuesto, incrementaba los costos.

La gráfica de control

El objetivo de una gráfica de control es determinar si el desempeño de un proceso se mantiene en un nivel aceptable de calidad. Se espera, desde luego, que cualquier proceso experimente una variabilidad natural, es decir, una variabilidad debida esencialmente a fuentes de variación poco importantes e incontrolables. Por otro lado, un proceso puede experimentar formas más severas de variabilidad en mediciones de desempeño fundamentales.

Estas fuentes de variabilidad pueden surgir de uno de varios tipos de "causas asignables" no aleatorias, como errores del operador o indicadores mal ajustados en una máquina. Un proceso que opera en dicho estado se denomina **fuera de control**. Se dice que un proceso que sólo experimenta variaciones aleatorias está en **control estadístico**. Desde luego, un proceso de producción exitoso puede operar en un estado de control durante un periodo largo. Se supone que durante este periodo el proceso elabora un producto aceptable. Sin embargo, podría ocurrir un "cambio" gradual o repentino que requiera detección.

El propósito de una gráfica de control es que funcione como un dispositivo para detectar el estado no aleatorio o fuera de control de un proceso. La gráfica de control suele adoptar la forma que se indica en la figura 17.1. Cuando ocurre un cambio en el proceso es importante detectarlo con rapidez, de manera que se pueda corregir el problema. Evidentemente, si el cambio no se detecta de inmediato, se producirán muchos artículos defectuosos o que no cumplen las especificaciones, lo cual dará como resultado un desperdicio significativo y un incremento en los costos.

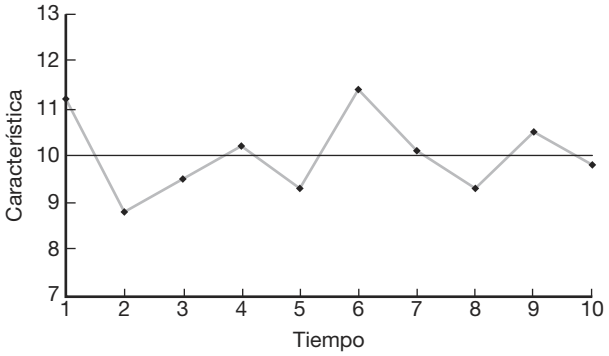


Figura 17.1: Gráfica de control típica.

Se deben considerar ciertos tipos de características de la calidad y se deben tomar muestras de las unidades del proceso a medida que pasa el tiempo. Digamos, por ejemplo, que la característica de un cojinete de motor es la circunferencia. La línea central representa el valor promedio de la característica cuando el proceso está bajo control. Los puntos que se indican en la figura representarían los resultados de, digamos, los promedios muestrales de tal característica, con muestras tomadas en diferentes momentos. Los límites de control superior e inferior se eligen de tal manera que se esperaría que, si el proceso está bajo control, todos los puntos muestrales queden cubiertos por estos límites. Como resultado, la forma general de los puntos graficados a lo largo del tiempo determina si se concluye que el proceso está bajo control. La evidencia de que está "dentro de control" se obtiene de un patrón aleatorio de puntos con todos los valores graficados dentro de los límites de control. Cuando un punto cae fuera de los límites de control, se considera como evidencia de que un proceso está fuera de control, en cuyo caso se sugiere una búsqueda para determinar la causa. Además, un patrón no aleatorio de puntos se debe considerar sospechoso y, evidentemente, un indicador de que es necesario investigar para encontrar la medida correctiva adecuada.

17.2 Naturaleza de los límites de control

Las ideas fundamentales en las que se basan las gráficas de control son similares en estructura a la prueba de hipótesis. Los límites de control se establecen para controlar la probabilidad de cometer el error de concluir que el proceso está fuera de control, cuando de hecho no lo está. Esto corresponde a la probabilidad de cometer un error tipo I si probáramos la hipótesis nula de que el proceso está bajo control. Por otro lado, debemos estar atentos al error del segundo tipo, es decir, el de no encontrar el proceso fuera de control cuando de hecho sí lo está (error tipo II). De esta manera, la elección de los límites de control es similar a la elección de una región crítica.

Como en el caso de la prueba de hipótesis, el tamaño de la muestra en cada punto es importante. La elección del tamaño de la muestra depende en gran medida de la sensibilidad o potencia de detección del estado fuera de control. En esta aplicación, el concepto de *potencia* es muy similar al de la situación de la prueba de hipótesis. Queda claro que cuanto más grande sea la muestra en cada periodo, más rápida será la detección de un proceso fuera de control. En cierto sentido los límites de control en realidad definen lo que el usuario considera como estar *bajo control.* En otras palabras, la amplitud dada por los límites de control debe depender en cierto sentido de la variabilidad del proceso. Como resultado, el cálculo de los límites de control dependerá de manera natural de los datos que se tomen de los resultados del proceso. De esta forma, cualquier aplicación del control de calidad debe comenzar con el cálculo de una muestra o conjunto de muestras preliminar, que establecerá tanto la línea central como los límites del control de calidad.

17.3 Objetivos de la gráfica de control

Un propósito evidente de la gráfica de control es la vigilancia del proceso, o sea determinar si es o no necesario realizar cambios. Además, la constante y sistemática obtención de datos a menudo permite a la administración evaluar la capacidad del proceso. Es evidente que, si una sola característica de desempeño es importante, el muestreo y la estimación continuos de la media y la desviación estándar de esa característica de desempeño ofrecen la actualización de lo que el proceso puede hacer en términos de desempeño promedio y variación aleatoria. Esto es valioso incluso cuando el proceso permanece bajo control durante periodos largos. La estructura sistemática y formal de la gráfica de control a menudo puede prevenir una reacción desmesurada ante cambios que representen sólo fluctuaciones aleatorias. Obviamente, en muchas situaciones los cambios realizados por una reacción desmesurada pueden crear graves problemas que son difíciles de resolver.

Las características de calidad de las gráficas de control por lo general caen en *dos* categorías: **variables** y **atributos**. Como resultado, los tipos de gráficas de control con frecuencia tienen las mismas clasificaciones. En el caso de la gráfica de los tipos de variables, la característica suele ser una medida sobre un continuo, como el diámetro o el peso. En el caso de la gráfica de atributos, lo que refleja la característica es si el producto individual se *ajusta a las especificaciones* (si está o no defectuoso). Las aplicaciones para estas dos situaciones distintas son evidentes.

En el caso de la gráfica de variables se debe ejercer control sobre la tendencia central y la variabilidad. Lo que a un analista de control de calidad le debe preocupar es si existe o no, *en promedio,* un cambio en los valores de la característica de desempeño. Además, siempre habrá interés por saber si algún cambio en las condiciones del proceso

provoca que disminuya la precisión, es decir, que aumente la variabilidad. Para manejar estos dos conceptos es esencial utilizar gráficas de control separadas. La tendencia central es controlada por la *gráfica* $\bar{X}$, donde las medias de muestras relativamente pequeñas se dibujan en la gráfica de control. La variabilidad alrededor de la media se controla mediante el *rango* en la muestra, o la *desviación estándar de la* muestra. En el caso de muestreo de atributos a menudo la cantidad que se grafica es la *proporción de artículos defectuosos* de una muestra. En la siguiente sección analizamos el desarrollo de gráficas de control para los tipos de variables de las características del desempeño.

17.4 Gráficas de control para variables

Un ejemplo es una forma relativamente sencilla de explicar los rudimentos de la gráfica $\bar{X}$ para variables. Suponga que en un proceso de fabricación de cierta parte de un motor se deben utilizar las gráficas de control de calidad. Suponga también que la media del proceso es $\mu = 50$ mm y que la desviación estándar es $\sigma = 0.01$ mm. Imagine que se toman muestras en grupos de 5 cada hora y que los valores de la *media muestral* $\bar{X}$ se registran y grafican como en la figura 17.2. Los límites para las gráficas $\bar{X}$ se basan en la desviación estándar de la variable aleatoria $\bar{X}$. Sabemos, a partir de lo expuesto en el capítulo 8, que para el promedio de observaciones independientes en una muestra de tamaño n,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

donde σ es la desviación estándar de una observación individual. Los límites de control están diseñados para dar como resultado una pequeña probabilidad de que un valor dado de $\bar{X}$ esté fuera de los límites dado que, en realidad, el proceso está bajo control, es decir, $\mu = 50$. Si recurrimos al teorema del límite central, tendremos que, en las condiciones en las que el proceso está controlado,

$$\bar{X} \sim N\left(50, \frac{0.01}{\sqrt{5}}\right).$$

Como resultado, $100(1 - \alpha)\%$ de los valores $\bar{X}$ cae dentro de los límites cuando el proceso está bajo control si utilizamos los límites

$$\text{LCI} = \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 - z_{\alpha/2}(0.0045), \qquad \text{LCS} = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 + z_{\alpha/2}(0.0045).$$

Aquí LCI y LCS representan el límite de control inferior y el límite de control superior, respectivamente. Con frecuencia las gráficas $\bar{X}$ se basan en límites denominados "tres-sigma", refiriéndonos, por supuesto, a $z_{\alpha/2} = 3$ y a límites que se convierten en

$$\mu \pm 3\frac{\sigma}{\sqrt{n}}.$$

En nuestro ejemplo, los límites superior e inferior son

$$\text{LCI} = 50 - 3(0.0045) = 49.9865, \qquad \text{LCS} = 50 + 3(0.0045) = 50.0135.$$

Por consiguiente, si vemos la estructura de los límites 3σ desde el punto de vista de la prueba de hipótesis para un punto muestral dado, encontraremos que hay una probabilidad

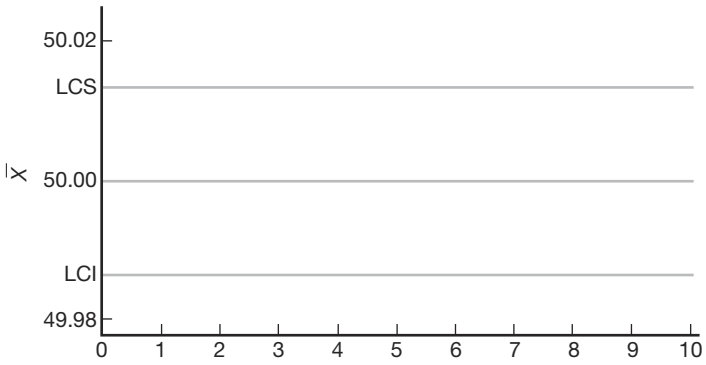


Figura 17.2: Los límites de control 3σ para el ejemplo de la parte del motor.

de 0.0026 de que el valor $\bar{X}$ caiga fuera de los límites de control, dado que el proceso está bajo control. Ésta es la probabilidad de que el analista determine *de manera errónea* que el proceso está fuera de control (véase la tabla A.3).

El ejemplo anterior no sólo ilustra la gráfica $\bar{X}$ para las variables, también proporciona al lector una idea general de la naturaleza de las gráficas de control. La línea central por lo general refleja el valor ideal de un parámetro importante. Los límites de control se establecen a partir del conocimiento de las propiedades de muestreo del estadístico que estima el parámetro en cuestión. Con mucha frecuencia implican un múltiplo de la desviación estándar del estadístico. Se ha generalizado el uso de límites 3σ. En el caso de la gráfica $\bar{X}$ que se presenta aquí, el teorema del límite central brinda al usuario una buena aproximación de la probabilidad de determinar de forma errónea que el proceso está fuera de control. En general, sin embargo, es probable que el usuario no confíe en la normalidad del estadístico sobre la línea central. Lo anterior podría dar como resultado que no se conozca la probabilidad exacta de cometer un "error tipo I". A pesar de esto se ha vuelto muy común utilizar los límites $k\sigma$. Aunque los límites 3σ se utilizan ampliamente, en ocasiones el usuario utilizará otro método. Cuando es importante detectar de forma rápida una situación fuera de control podría ser apropiado utilizar un múltiplo menor de σ. Si se toman en cuenta los costos de producción, cabe señalar que permitir que un proceso continúe funcionando fuera de control, incluso por periodos cortos, puede resultar más costoso que invertir en la investigación y corrección de las causas de la pérdida del control en el proceso. En este caso es evidente que los límites apropiados son los límites de control que son más estrictos que los límites 3σ.

Subgrupos racionales

Los valores de la muestra que se utilizan para el control de calidad se dividen en subgrupos, en los que una *muestra* representa un subgrupo. Como antes indicamos, el orden en el tiempo de producción es en realidad una base natural para la selección de los subgrupos. Podríamos considerar el esfuerzo de control de calidad de manera muy simple como 1) muestreo, 2) detección de un estado fuera de control y 3) búsqueda de las causas atribuibles que puedan ocurrir con el tiempo. Tal vez parezca que la selección de la base para estos grupos muestrales es muy sencilla, pero la elección de estos subgrupos de información muestral podría tener un efecto importante en el éxito del programa de control de calidad. Estos subgrupos con frecuencia se denominan **subgrupos racionales**. En

general, si el analista está interesado en detectar un *cambio de ubicación,* se considera que los subgrupos se deben elegir de manera que la variabilidad dentro del subgrupo sea pequeña, y de manera que haya mayores posibilidades de detectar las causas atribuibles, si se presentaran. Así, deseamos elegir los subgrupos de forma que se maximice la variabilidad entre subgrupos. Por ejemplo, un método razonable es elegir unidades en un subgrupo que se producen de forma cercana en el tiempo. Por otro lado, las gráficas de control a menudo se utilizan para controlar la variabilidad, en cuyo caso el estadístico de desempeño es la *variabilidad dentro de la muestra.* Por consiguiente, es más importante elegir los subgrupos racionales para maximizar la variabilidad dentro de la muestra. En este caso las observaciones en los subgrupos se deberían comportar más como una muestra aleatoria y la variabilidad dentro de las muestras necesita ser una descripción de la variabilidad del proceso.

Es importante señalar que las gráficas de control sobre la variabilidad se deben establecer antes de construir gráficas sobre el centro de ubicación (digamos, gráficas $\bar{X}$). Cualquier gráfica de control sobre el centro de ubicación en realidad dependerá de la variabilidad. Por ejemplo, vimos un ejemplo de la gráfica de tendencia central y ésta depende de σ. En las secciones que siguen se analizará un estimado de σ a partir de los datos.

Gráfica $\bar{X}$ con parámetros estimados

Con anterioridad ilustramos las nociones de la gráfica $\bar{X}$ que usa el teorema del límite central y emplea valores *conocidos* de la media y desviación estándar del proceso. Como al principio se indicó, se utilizan los límites de control

$$\text{LCI} = \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \text{LCS} = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

y un valor $\bar{X}$ que cae fuera de estos límites se considera evidencia de un cambio en la media μ, y, por lo tanto, de la posibilidad de que el proceso esté fuera de control.

En muchas situaciones prácticas no es razonable suponer que conocemos μ y σ. Como resultado, se deben proporcionar estimados de los datos que se obtienen cuando el proceso está bajo control. Por lo general los estimados se determinan durante un periodo en el que se reúne *información antecedente* o *de inicio.* Se elige una base para subgrupos racionales y se reúnen los datos con muestras de tamaño n en cada subgrupo. Los tamaños de la muestra por lo general son pequeños, digamos, 4, 5 o 6, y se toman k muestras, con k al menos igual a 20. Durante este periodo, en el que se supone que el proceso está bajo control, el usuario establece los estimados de μ y σ en los que se basa la gráfica de control. La información importante reunida durante este periodo incluye las medias muestrales en el subgrupo, la media general y el rango de la muestra en cada subgrupo. En los siguientes párrafos señalaremos cómo se utiliza esta información para producir la gráfica de control.

Una parte de la información muestral de estas k muestras toma la forma $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$, donde la variable aleatoria $\bar{X}_i$ es el promedio de los valores en la i-ésima muestra. Evidentemente, el promedio global es la variable aleatoria

$$\bar{\bar{X}} = \frac{1}{k}\sum_{i=1}^{k}\bar{X}_i.$$

Éste es el estimador adecuado de la media del proceso y, por consiguiente, es la línea central en la gráfica de control $\bar{X}$. En aplicaciones de control de calidad a menudo es

conveniente estimar σ a partir de la información relacionada con los *rangos* en las muestras, en vez de las desviaciones estándar muestrales. Definamos

$$R_i = X_{\text{máx},\,i} - X_{\text{mín},\,i}$$

como el rango para los datos en la i-ésima muestra. Aquí $X_{\text{máx},i}$ y $X_{\text{mín},i}$ son, respectivamente, la observación más grande y la más pequeña en la muestra. El estimado apropiado de σ es una función del rango promedio

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i.$$

Un estimado de σ, digamos $\hat{\sigma}$, se obtiene mediante

$$\hat{\sigma} = \frac{\bar{R}}{d_2},$$

donde d_2 es una constante que depende del tamaño de la muestra. Los valores de d_2 se muestran en la tabla A.22.

El uso del rango para producir un estimado de σ tiene sus raíces en aplicaciones similares a la del control de calidad, en particular debido a que, en la época en que aún era muy difícil lograr cálculos precisos, el rango era muy fácil de calcular en comparación con otros estimados de variabilidad. La suposición de normalidad de las observaciones individuales está implícita en la gráfica $\bar{X}$. Por supuesto, la existencia del teorema del límite central es ciertamente útil a este respecto. Bajo la suposición de normalidad, usamos una variable aleatoria llamada rango relativo dada por

$$W = \frac{R}{\sigma}.$$

De la cual resulta que los momentos de W son funciones simples del tamaño de la muestra n (véase la referencia de Montgomery, 2000b, en la bibliografía). El valor esperado de W a menudo se denomina d_2. Así, al tomar el valor esperado de la W anterior,

$$\frac{E(R)}{\sigma} = d_2,$$

la cual facilita la comprensión del fundamento para el estimado $\hat{\sigma} = \bar{R}/d_2$. Se sabe bien que el método del rango produce un estimador eficiente de σ en muestras hasta cierto punto pequeñas. Esto hace que el estimador sea particularmente atractivo en aplicaciones de control de calidad, ya que los tamaños de la muestra en los subgrupos por lo general son pequeños. El uso del método del rango para la estimación de σ tiene como resultado gráficas de control con los siguientes parámetros:

$$\text{LCS} = \bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}, \qquad \text{línea central} = \bar{\bar{X}}, \qquad \text{LCI} = \bar{\bar{X}} - \frac{3\bar{R}}{d_2\sqrt{n}}.$$

Al definir la cantidad

$$A_2 = \frac{3}{d_2\sqrt{n}},$$

tenemos que

$$\text{LCS} = \bar{\bar{X}} + A_2 \bar{R}, \qquad \text{LCI} = \bar{\bar{X}} - A_2 \bar{R}.$$

Para simplificar la estructura el usuario de las gráficas $\bar{X}$ a menudo encuentra valores tabulados de A_2. En la tabla A.22 se incluyen valores de A_2 para varios tamaños de la muestra.

Gráficas *R* para control de variación

Hasta aquí todos los ejemplos y detalles tuvieron que ver con el intento del analista de control de calidad de detectar condiciones fuera de control producidas por un *cambio en la media.* Los límites de control se basan en la distribución de la variable aleatoria $\bar{X}$ y dependen de la suposición de normalidad de las observaciones individuales. Es importante que el control se aplique tanto a la variabilidad como al centro de ubicación. De hecho, muchos expertos consideran que el control de la variabilidad de la característica del desempeño es más importante y que es necesario establecerlo antes de considerar el centro de ubicación. La variabilidad del proceso se puede controlar usando *gráficas del rango muestral.* Una gráfica de los rangos muestrales a lo largo del tiempo se denomina **gráfica *R***. Se puede utilizar la misma estructura general, como en el caso de la gráfica $\bar{X}$, donde $\bar{R}$ es la *línea central* y los límites de control dependen de que se estime la desviación estándar de la variable aleatoria *R.* Por lo tanto, como en el caso de la gráfica $\bar{X}$, se establecen límites 3σ donde "3σ" implica $3\sigma_R$. La cantidad σ_R se debe estimar a partir de los datos, tal como se estima $\sigma_{\bar{X}}$.

El estimado de σ_R, la desviación estándar, también se basa en la distribución del rango relativo

$$W = \frac{R}{\sigma}.$$

La desviación estándar de W es una función conocida del tamaño de la muestra y por lo general se denota por d_3. Esto da como resultado,

$$\sigma_R = \sigma d_3.$$

Ahora podemos reemplazar σ por $\hat{\sigma} = \bar{R}/d_2$, y de esta forma el estimador de σ_R es

$$\hat{\sigma}_R = \frac{\bar{R} d_3}{d_2}.$$

Por consiguiente, las cantidades que definen la gráfica *R* son

$$\text{LCS} = \bar{R} D_4, \qquad \text{línea central} = \bar{R}, \qquad \text{LCI} = \bar{R} D_3,$$

donde las constantes D_4 y D_3 (que dependen sólo de n) son

$$D_4 = 1 + 3\frac{d_3}{d_2}, \qquad D_3 = 1 - 3\frac{d_3}{d_2}.$$

Las constantes D_4 y D_3 se encuentran tabuladas en la tabla A.22.

Gráficas $\bar{X}$ y R para variables

Se controla un proceso de fabricación de partes componentes para misiles, donde la característica de desempeño es la resistencia a la tensión, en libras por pulgada cuadrada. Se toman muestras de tamaño 5 cada hora y se reportan 25 muestras. Los datos se muestran en la tabla 17.1.

Tabla 17.1: Información muestral de los datos de resistencia a la tensión

Número de muestra	Observaciones					$\bar{X}_i$	R_i
1	1515	1518	1512	1498	1511	1510.8	20
2	1504	1511	1507	1499	1502	1504.6	12
3	1517	1513	1504	1521	1520	1515.0	17
4	1497	1503	1510	1508	1502	1504.0	13
5	1507	1502	1497	1509	1512	1505.4	15
6	1519	1522	1523	1517	1511	1518.4	12
7	1498	1497	1507	1511	1508	1504.2	14
8	1511	1518	1507	1503	1509	1509.6	15
9	1506	1503	1498	1508	1506	1504.2	10
10	1503	1506	1511	1501	1500	1504.2	11
11	1499	1503	1507	1503	1501	1502.6	8
12	1507	1503	1502	1500	1501	1502.6	7
13	1500	1506	1501	1498	1507	1502.4	9
14	1501	1509	1503	1508	1503	1504.8	8
15	1507	1508	1502	1509	1501	1505.4	8
16	1511	1509	1503	1510	1507	1508.0	8
17	1508	1511	1513	1509	1506	1509.4	7
18	1508	1509	1512	1515	1519	1512.6	11
19	1520	1517	1519	1522	1516	1518.8	6
20	1506	1511	1517	1516	1508	1511.6	11
21	1500	1498	1503	1504	1508	1502.6	10
22	1511	1514	1509	1508	1506	1509.6	8
23	1505	1508	1500	1509	1503	1505.0	9
24	1501	1498	1505	1502	1505	1502.2	7
25	1509	1511	1507	1500	1499	1505.2	12

Como antes indicamos, es importante comenzar por establecer las condiciones de variabilidad "bajo control". La línea central calculada para la gráfica R es

$$\bar{R} = \frac{1}{25}\sum_{i=1}^{25} R_i = 10.72.$$

En la tabla A.22 encontramos que para $n = 5$, $D_3 = 0$ y $D_4 = 2.114$. Como resultado, los límites de control para la gráfica R son

$$\text{LCI} = \bar{R}D_3 = (10.72)(0) = 0,$$
$$\text{LCS} = \bar{R}D_4 = (10.72)(2.114) = 22.6621.$$

En la figura 17.3 se muestra la gráfica *R*. Ninguno de los rangos graficados cae fuera de los límites de control. Como resultado, no hay nada que indique la existencia de una situación fuera de control.

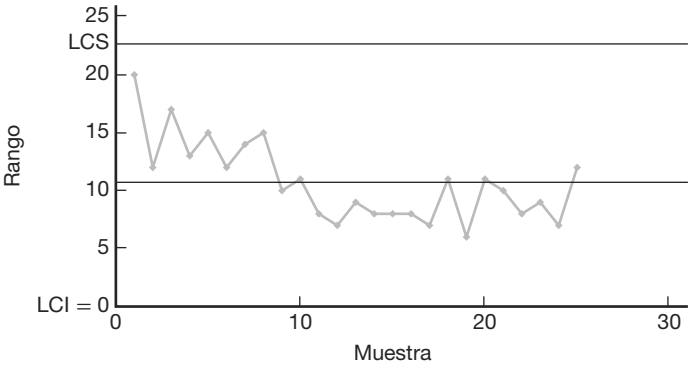


Figura 17.3: Gráfica *R* para el ejemplo de resistencia a la tensión.

Ahora se puede construir la gráfica $\bar{X}$ para las lecturas de la resistencia a la tensión. La línea central es

$$\bar{\bar{X}} = \frac{1}{25}\sum_{i=1}^{25}\bar{X}_i = 1507.328.$$

En la tabla A.22 encontramos que, para muestras de tamaño 5, $A_2 = 0.577$. De esta forma, los límites de control son

$$\text{LCS} = \bar{\bar{X}} + A_2\bar{R} = 1507.328 + (0.577)(10.72) = 1513.5134,$$
$$\text{LCI} = \bar{\bar{X}} - A_2\bar{R} = 1507.328 - (0.577)(10.72) = 1501.1426.$$

En la figura 17.4 se muestra la gráfica $\bar{X}$. Como el lector puede observar, tres valores caen fuera de los límites de control, lo cual es una señal de que no se deberían usar los límites de control de $\bar{X}$ para el control de calidad de la línea.

Más comentarios acerca de las gráficas de control para variables

Un proceso podría parecer estar bajo control y, de hecho, permanecer así durante un periodo largo. ¿Esto significaría necesariamente que el proceso está funcionando de manera exitosa? Un proceso que opera *bajo control* es simplemente aquel en el que la media y la variabilidad del proceso permanecen estables, indicando, aparentemente, que no han ocurrido cambios graves. "Bajo control" implica que el proceso permanece consistente con variabilidad natural. Las gráficas de control de calidad pueden verse como un método en el que la variabilidad natural inherente rige la amplitud de los límites de control. Sin embargo, no determinan hasta qué punto un proceso bajo control satisface las *especificaciones* predeterminadas que requiere el proceso. Las especificaciones son límites que establece el consumidor. Si la variabilidad natural del proceso actual es mayor

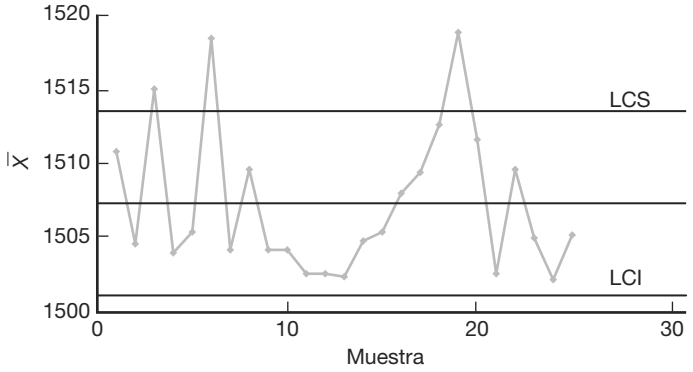


Figura 17.4: Gráfica $\bar{X}$ para el ejemplo de resistencia a la tensión.

que la que determinan las especificaciones, aunque el proceso permanezca estable y esté bajo control, con demasiada frecuencia producirá artículos que no cumplirán las especificaciones.

Aludimos a la suposición de normalidad para las observaciones individuales en una gráfica de control de variables. Para la gráfica $\bar{X}$, si las observaciones individuales son normales, el estadístico $\bar{X}$ es normal. Como resultado, el analista de control de calidad en este caso tiene control sobre la probabilidad de un error tipo I. Si las X individuales no son normales, $\bar{X}$ es aproximadamente normal, por lo tanto, existe un control aproximado sobre la probabilidad de un error tipo I para el caso en el que se conoce σ. Sin embargo, utilizar o no el método del rango para estimar la desviación estándar también depende de la suposición de normalidad. Estudios respecto a la robustez de la gráfica $\bar{X}$ para desviaciones de la normalidad indican que, para las muestras de tamaño $k \geq 4$, la gráfica $\bar{X}$ da como resultado un riesgo α cercano al anunciado (véase el trabajo de Montgomery, 2000b, y Schilling y Nelson, 1976, en la bibliografía). Indicamos antes que la aproximación $\pm k\sigma_R$ a la gráfica R es una cuestión de conveniencia y tradición. Incluso si la distribución de observaciones individuales es normal, la distribución de R no es normal. De hecho, la distribución de R no es ni siquiera simétrica. Los límites de control simétricos de $\pm k\sigma_R$ sólo proporcionan una aproximación al riesgo α y, en algunos casos, la aproximación no es particularmente buena.

Elección del tamaño de la muestra (función característica de operación) en el caso de la gráfica $\bar{X}$

Los científicos e ingenieros que manejan el control de calidad a menudo se refieren a los factores que influyen en el *diseño de la gráfica de control.* Los componentes que determinan el diseño de la gráfica incluyen el tamaño de la muestra que se toma en cada subgrupo, la amplitud de los límites de control y la frecuencia del muestreo. Todos estos factores dependen en gran medida de consideraciones económicas y prácticas. La frecuencia de muestreo evidentemente depende del costo del muestreo y del costo en el que se incurre si el proceso continúa fuera de control durante un periodo largo. Estos mismos factores afectan la amplitud de la región "bajo control". El costo asociado con la investigación y la búsqueda de las causas atribuibles de la pérdida de control repercute en

la amplitud de la región y en la frecuencia de muestreo. Se ha puesto mucha atención en el diseño óptimo de gráficas de control, por lo que aquí no se darán mayores detalles. Se remite al lector al trabajo de Montgomery (2000b), que se cita en la bibliografía, para un excelente recuento histórico de gran parte de esta investigación.

La elección del tamaño de la muestra y la frecuencia de muestreo implican equilibrar los recursos disponibles para estos dos esfuerzos. En muchos casos es probable que el analista necesite hacer cambios en la estrategia hasta lograr el equilibrio adecuado. El analista siempre debe estar consciente de que, si el costo de producción de artículos no adecuados es grande, la estrategia adecuada sería una alta frecuencia de muestreo con un tamaño de la muestra relativamente pequeño.

Al elegir el tamaño de una muestra hay que tomar en cuenta muchos factores. En la ilustración y el análisis enfatizamos el uso de $n = 4$, 5 o 6. Estos valores se consideran relativamente pequeños para problemas generales en inferencia estadística, pero serían tamaños de muestra apropiados para el control de calidad. Una justificación, por supuesto, es que el control de calidad es un proceso continuo y los resultados producidos por una muestra o un conjunto de unidades serán seguidos por resultados de muchas más. Así, el tamaño de la muestra "eficaz" de todo el esfuerzo de control de calidad es muchas veces mayor que el tamaño que se utiliza en un subgrupo. Por lo general se considera más efectivo *tomar muestras frecuentemente* con un tamaño muestral pequeño.

El analista puede utilizar el concepto de potencia de una prueba para obtener información de la eficacia del tamaño de la muestra elegido. Esto es especialmente importante, ya que por lo general se utilizan muestras de tamaño pequeño en cada subgrupo. Remítase a los capítulos 10 y 13 para un análisis de la potencia de pruebas formales sobre las medias y el análisis de varianza. Aunque en el control de calidad en realidad no se realizan pruebas formales de hipótesis, se puede tratar la información como si la estrategia en cada subgrupo fuera la de probar una hipótesis, ya sea sobre la media de la población μ o sobre la desviación estándar σ. Es de interés la *probabilidad de detectar* una condición fuera de control para una muestra dada y, quizá más importante, el número esperado de corridas requeridas para detectarla. La probabilidad de detectar una condición fuera de control específica corresponde a la potencia de una prueba. No es nuestra intención demostrar el desarrollo de la potencia para todos los tipos de gráficas de control que aquí se presentan, más bien, lo que deseamos es mostrar el desarrollo de la gráfica $\bar{X}$ y presentar los resultados de potencia para la gráfica R.

Considere la gráfica $\bar{X}$ cuando se conoce el valor de σ. Suponga que el estado bajo control tiene $\mu = \mu_0$. Un estudio del papel que desempeña el tamaño de la muestra del subgrupo equivale a investigar el riesgo β, es decir, la probabilidad de que un valor $\bar{X}$ permanezca dentro de los límites de control cuando realmente ha ocurrido un cambio en la media. Suponga que la forma que toma el cambio es

$$\mu = \mu_0 + r\sigma.$$

De nuevo, al utilizar la normalidad de $\bar{X}$ tenemos

$$\beta = P(\text{LCI} \leq \bar{X} \leq \text{LCS} \mid \mu = \mu_0 + r\sigma).$$

Para el caso de límites $k\sigma$,

$$\text{LCI} = \mu_0 - \frac{k\sigma}{\sqrt{n}} \quad \text{y} \quad \text{LCS} = \mu_0 + \frac{k\sigma}{\sqrt{n}}.$$

Como resultado, si denotamos con Z la variable aleatoria normal estándar

$$\begin{aligned}\beta &= P\left[Z < \left(\frac{\mu_0 + k\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right)\right] - P\left[Z < \left(\frac{\mu_0 - k\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right)\right] \\ &= P\left\{Z < \left[\frac{\mu_0 + k\sigma/\sqrt{n} - (\mu + r\sigma)}{\sigma/\sqrt{n}}\right]\right\} - P\left\{Z < \left[\frac{\mu_0 - k\sigma/\sqrt{n} - (\mu + r\sigma)}{\sigma/\sqrt{n}}\right]\right\} \\ &= P(Z < k - r\sqrt{n}) - P(Z < -k - r\sqrt{n}).\end{aligned}$$

Observe el papel que desempeñan *n, r* y *k* en la expresión para el riesgo β. La probabilidad de no detectar un cambio específico, como se esperaba, aumenta claramente con un incremento en *k*. β disminuye con un aumento en *r,* la magnitud del cambio, y disminuye con un incremento en el tamaño de la muestra, *n.*

Se debería enfatizar que la expresión anterior da como resultado el riesgo β (probabilidad de un error tipo II) para el caso de una *sola muestra.* Por ejemplo, suponga que, en el caso de una muestra de tamaño 4, ocurre un cambio de σ en la media. La probabilidad de detectar el cambio (potencia) *en la primera muestra después del cambio* es, suponiendo límites 3σ:

$$1 - \beta = 1 - [P(Z < 1) - P(Z < -5)] = 0.1587.$$

Por otro lado, la probabilidad de detectar un cambio de 2σ es

$$1 - \beta = 1 - [P(Z < -1) - P(Z < -7)] = 0.8413.$$

Los resultados anteriores ilustran una muy modesta probabilidad de detectar un cambio de magnitud σ y una alta probabilidad de detectar un cambio de magnitud 2σ. En la figura 17.5 se observa la imagen completa de cómo se desempeñan los límites de control 3σ para la gráfica $\bar{X}$ que aquí se describe. En lugar de graficar las funciones de potencia se presenta una gráfica de β contra *r,* donde el cambio en la media tiene una magnitud $r\sigma$. Por supuesto, los tamaños de la muestra de $n = 4, 5, 6$ dan como resultado una pequeña probabilidad de detectar un cambio de 1.0σ o incluso 1.5σ en la primera muestra después del cambio.

Pero si el muestreo se realiza con frecuencia, la probabilidad podría no ser tan importante como el número promedio o esperado de corridas que se requiere antes de detectar un cambio. Una detección rápida es importante y ciertamente posible, aunque no hay muchas probabilidades de lograrlo en la primera muestra. Resulta que las gráficas $\bar{X}$ con estas muestras pequeñas conducirán a una detección relativamente rápida. Si β es la probabilidad de no detectar un cambio en la primera muestra después del cambio, entonces la probabilidad de detectarlo en la muestra *s*-ésima después de que ocurre es, suponiendo que las muestras son independientes:

$$P_s = (1-\beta)\beta^{s-1}.$$

El lector debe reconocer que ésta es una aplicación de la distribución geométrica. El valor promedio o esperado del número de muestras que se requieren para la detección es

$$\sum_{s=1}^{\infty} s\beta^{s-1}(1-\beta) = \frac{1}{1-\beta}.$$

Por consiguiente, el número esperado de muestras que se requieren para detectar el cambio en la media es el *recíproco de la potencia,* es decir, la probabilidad de detección en la primera muestra después del cambio.

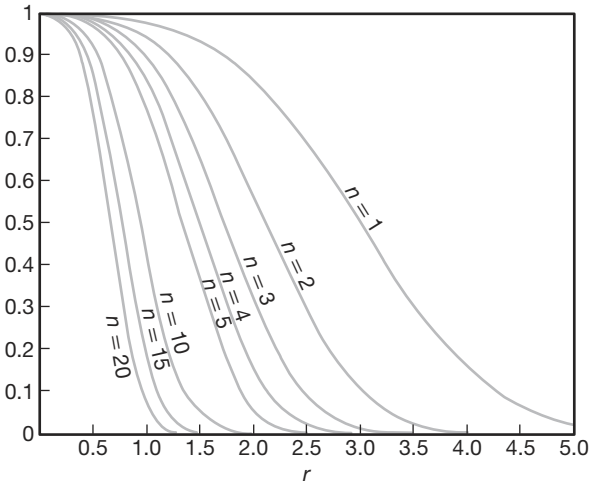


Figura 17.5: Curvas características de operación para la gráfica $\bar{X}$ con límites 3σ. Aquí, β es el error de probabilidad tipo II en la primera muestra después de que ocurre un cambio en la media de $r\sigma$.

Ejemplo 17.1: En cierto esfuerzo por controlar la calidad es importante que el analista detecte con rapidez los cambios en la media de $\pm\sigma$ utilizando una gráfica de control 3σ con una muestra de tamaño $n = 4$. El número esperado de muestras que se requieren después del cambio para detectar el estado fuera de control podría ser útil en la evaluación del procedimiento de control de calidad.

En la figura 17.5, para $n = 4$ y $r = 1$, se puede ver que $\beta \approx 0.84$. Si utilizamos s para denotar el número de muestras que se requieren para detectar el cambio, la media de s es

$$E(s) = \frac{1}{1-\beta} = \frac{1}{0.16} = 6.25.$$

De esta manera, se requieren siete subgrupos, en promedio, antes de detectar un cambio de $\pm\sigma$. ∎

Elección del tamaño de la muestra para la gráfica *R*

La curva CO de la gráfica *R* se muestra en la figura 17.6. Como la gráfica *R* se utiliza para controlar la desviación estándar del proceso, y la desviación estándar después de que el proceso se sale de control, el riesgo β se grafica como una función de la desviación estándar bajo control, σ_0. La última desviación estándar se denotará con σ_1. Sea

$$\lambda = \frac{\sigma_1}{\sigma_0}.$$

Para varios tamaños muestrales se grafica β contra λ.

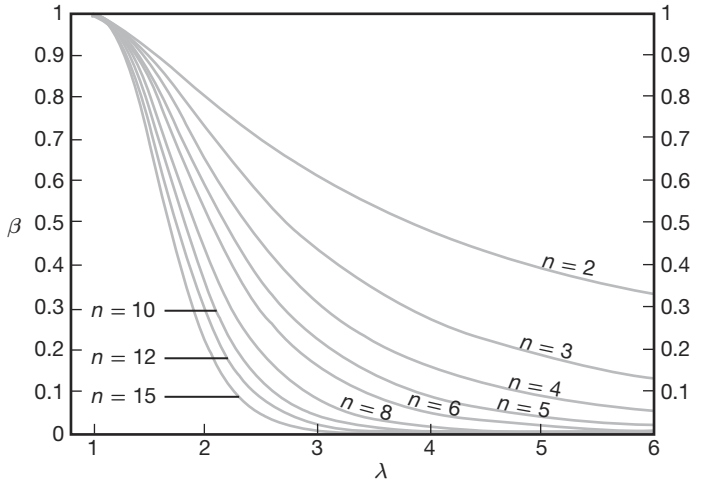


Figura 17.6: Curva característica operativa para las gráficas R con límites de 3σ.

Gráficas $\bar{X}$ y S para variables

Para el estudiante de estadística es natural anticipar el uso de la varianza muestral en la gráfica $\bar{X}$ y en una gráfica para el control de la variabilidad. El rango es un estimador eficiente de σ, pero esta eficiencia disminuye a medida que aumenta el tamaño de la muestra. Para una n tan grande como 10 se debe utilizar el tan conocido estadístico

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

en la gráfica de control, tanto para la media como para la variabilidad. El lector debe recordar que en el capítulo 9 se expuso que S^2 es un estimador no sesgado de σ^2, pero que S no es no sesgado para σ. Para evitar sesgos se acostumbra corregir S en las aplicaciones de la gráfica de control. Sabemos, en general, que

$$E(S) \neq \sigma.$$

En el caso en que las X_i sean independientes y estén distribuidas de forma normal con media μ y varianza σ^2,

$$E(S) = c_4\sigma, \qquad \text{donde} \qquad c_4 = \left(\frac{2}{n-1}\right)^{1/2}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$$

y $\Gamma(\cdot)$ se refiere a la función gamma (véase el capítulo 6). Por ejemplo, para $n = 5$, $c_4 = (3/8)\sqrt{2\pi}$. Además, la varianza del estimador S es

$$\text{Var}(S) = \sigma^2(1 - c_4^2).$$

Establecimos las propiedades de S que nos permitirán escribir límites de control para $\bar{X}$ y S. Para construir una estructura adecuada comenzamos por suponer que conocemos σ. Después presentamos la estimación de σ a partir de un conjunto de muestras preliminar.

Al graficar el estadístico S, los parámetros evidentes de la gráfica de control son

$$\text{LCS} = c_4\sigma + 3\sigma\sqrt{1 - c_4^2}, \quad \text{línea central} = c_4\sigma, \quad \text{LCI} = c_4\sigma - 3\sigma\sqrt{1 - c_4^2}.$$

Como de costumbre, los límites de control se definen de manera más sucinta utilizando constantes tabuladas. Sean

$$B_5 = c_4 - 3\sqrt{1 - c_4^2}, \qquad B_6 = c_4 + 3\sqrt{1 - c_4^2},$$

entonces, tenemos

$$\text{LCS} = B_6\sigma, \quad \text{línea central} = c_4\sigma, \qquad \text{LCI} = B_5\sigma.$$

En la tabla A.22 se encuentran tabulados los valores de B_5 y B_6 para varios tamaños muestrales.

Ahora, por supuesto, los límites de control anteriores sirven como base para el desarrollo de los parámetros de control de calidad en la situación que con más frecuencia se observa en la práctica, a saber, en la que se desconoce σ. Debemos suponer una vez más que para producir un estimado de σ durante lo que se supone es un periodo "bajo control" se toma un conjunto de *muestras base* o muestras preliminares. Las desviaciones estándar de las muestras $S_1, S_2, \ldots, S_m$ se obtienen a partir de muestras que son, cada una, de tamaño n. A menudo se utiliza un estimador no sesgado del tipo

$$\frac{\bar{S}}{c_4} = \left(\frac{1}{m}\sum_{i=1}^{m} S_i\right) \Big/ c_4$$

para σ. Aquí, desde luego, $\bar{S}$, el valor promedio de la desviación estándar muestral en la muestra preliminar, es la línea central lógica en la gráfica de control para el control de la variabilidad. Los límites de control superior e inferior son estimadores no sesgados de los límites de control adecuados para el caso en el que se conoce σ. Como

$$E\left(\frac{\bar{S}}{c_4}\right) = \sigma,$$

el estadístico $\bar{S}$ es una línea central apropiada (como un estimador no sesgado de $c_4\sigma$) y las cantidades

$$\bar{S} - 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2} \qquad \text{y} \qquad \bar{S} + 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2}$$

son los límites de control 3σ inferior y superior apropiados, respectivamente. Como resultado, la línea central y los límites para la gráfica S de control de variabilidad son

$$\text{LCI} = B_3\bar{S}, \qquad \text{línea central} = \bar{S}, \qquad \text{LCS} = B_4\bar{S},$$

donde

$$B_3 = 1 - \frac{3}{c_4}\sqrt{1 - c_4^2}, \qquad B_4 = 1 + \frac{3}{c_4}\sqrt{1 - c_4^2}.$$

las constantes B_3 y B_4 aparecen en la tabla A.22.

Ahora podemos escribir los parámetros de la gráfica $\bar{X}$ correspondiente que implican el uso de la desviación estándar muestral. Supongamos que podemos disponer de S y $\bar{X}$ de la muestra base preliminar. La línea central continúa siendo $\bar{\bar{X}}$ y los límites 3σ son simplemente de la forma $\bar{\bar{X}} \pm 3\hat{\sigma}/\sqrt{n}$, donde $\hat{\sigma}$ es un estimador insesgado. Simplemente proporcionamos $\bar{S}/c_4$ como un estimador de σ, y de esta forma tenemos

$$\text{LCI} = \bar{\bar{X}} - A_3\,\bar{S}, \qquad \text{línea central} = \bar{\bar{X}}, \qquad \text{LCS} = \bar{\bar{X}} + A_3\,\bar{S},$$

donde

$$A_3 = \frac{3}{c_4\sqrt{n}}.$$

En la tabla A.22 aparece la constante A_3 para varios tamaños de la muestra.

Ejemplo 17.2: Se producen contenedores mediante un proceso en el que el volumen de éstos es sometido a un control de calidad. Se utilizaron 25 muestras de tamaño 5 para establecer los parámetros de control de calidad. En la tabla 17.2 se documenta la información de estas muestras.

En la tabla A.22 se observa que $B_3 = 0$, $B_4 = 2.089$ y $A_3 = 1.427$. Como resultado, los límites de control para $\bar{X}$ son dados por

$$\text{LCS} = \bar{\bar{X}} + A_3\,\bar{S} = 62.3771, \qquad \text{LCI} = \bar{\bar{X}} - A_3\,\bar{S} = 62.2741,$$

y los límites de control para la gráfica S son

$$\text{LCI} = B_3\,\bar{S} = 0, \qquad \text{LCS} = B_4\,\bar{S} = 0.0754.$$

Las figuras 17.7 y 17.8 muestran las gráficas de control para este ejemplo, $\bar{X}$ y S, respectivamente. En las gráficas se representa la información de las 25 muestras en el conjunto de datos preliminar. Al parecer, el control se establece después de las primeras muestras. ▟

17.5 Gráficas de control para atributos

Como indicamos al principio de este capítulo, muchas aplicaciones industriales de control de calidad requieren que la característica de calidad indique sólo que el artículo "se ajusta". En otras palabras, no hay una medición continua que sea crucial para el desempeño del artículo. Una ilustración evidente de este tipo de muestreo, denominado **muestreo por atributos**, es el desempeño de una bombilla que funciona o no de manera satisfactoria. El artículo **está o no defectuoso**. Las piezas metálicas fabricadas pueden tener deformaciones; los contenedores de una línea de producción pueden tener fugas. En ambos casos un artículo defectuoso impide su uso por parte del consumidor. La gráfica de control estándar para esta situación es la gráfica p, o *gráfica para la fracción de defectuosos.* Como se podría esperar, la distribución de probabilidad que interviene es la distribución binomial. Se remite al lector al capítulo 5 para información básica de la distribución binomial.

Tabla 17.2: Volumen de contenedores para 25 muestras en una muestra preliminar (en centímetros cúbicos)

Muestra	Observaciones					$\bar{X}_i$	S_i
1	62.255	62.301	62.289	62.189	62.311	62.269	0.0495
2	62.187	62.225	62.337	62.297	62.307	62.271	0.0622
3	62.421	62.377	62.257	62.295	62.222	62.314	0.0829
4	62.301	62.315	62.293	62.317	62.409	62.327	0.0469
5	62.400	62.375	62.295	62.272	62.372	62.343	0.0558
6	62.372	62.275	62.315	62.372	62.302	62.327	0.0434
7	62.297	62.303	62.337	62.392	62.344	62.335	0.0381
8	62.325	62.362	62.351	62.371	62.397	62.361	0.0264
9	62.327	62.297	62.318	62.342	62.318	62.320	0.0163
10	62.297	62.325	62.303	62.307	62.333	62.313	0.0153
11	62.315	62.366	62.308	62.318	62.319	62.325	0.0232
12	62.297	62.322	62.344	62.342	62.313	62.324	0.0198
13	62.375	62.287	62.362	62.319	62.382	62.345	0.0406
14	62.317	62.321	62.297	62.372	62.319	62.325	0.0279
15	62.299	62.307	62.383	62.341	62.394	62.345	0.0431
16	62.308	62.319	62.344	62.319	62.378	62.334	0.0281
17	62.319	62.357	62.277	62.315	62.295	62.313	0.0300
18	62.333	62.362	62.292	62.327	62.314	62.326	0.0257
19	62.313	62.387	62.315	62.318	62.341	62.335	0.0313
20	62.375	62.321	62.354	62.342	62.375	62.353	0.0230
21	62.399	62.308	62.292	62.372	62.299	62.334	0.0483
22	62.309	62.403	62.318	62.295	62.317	62.328	0.0427
23	62.293	62.293	62.342	62.315	62.349	62.318	0.0264
24	62.388	62.308	62.315	62.392	62.303	62.341	0.0448
25	62.324	62.318	62.315	62.295	62.319	62.314	0.0111
						$\bar{\bar{X}} = 62.3256$	
						$\bar{S} = 0.0361$	

Gráfica *p* para la fracción de artículos defectuosos

Cualquier artículo fabricado puede tener varias características que son importantes y deben ser examinadas por un inspector. Sin embargo, todo el procedimiento se enfoca aquí en una sola característica. Suponga que para todos los artículos la probabilidad de encontrar uno defectuoso es p, y que todos los artículos se producen de forma independiente. Entonces, en una muestra aleatoria de n artículos producidos, con X como el número de artículos defectuosos, tenemos

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

Como se podría suponer, la media y varianza de la variable aleatoria binomial desempeñarán un papel importante en el desarrollo de la gráfica de control. El lector debería recordar que

$$E(X) = np \quad \text{y} \quad \text{Var}(X) = np(1-p).$$

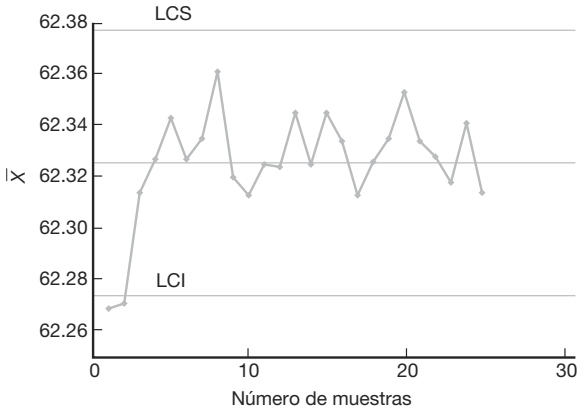


Figura 17.7: Gráfica $\bar{X}$ con límites de control establecidos con los datos del ejemplo 17.2.

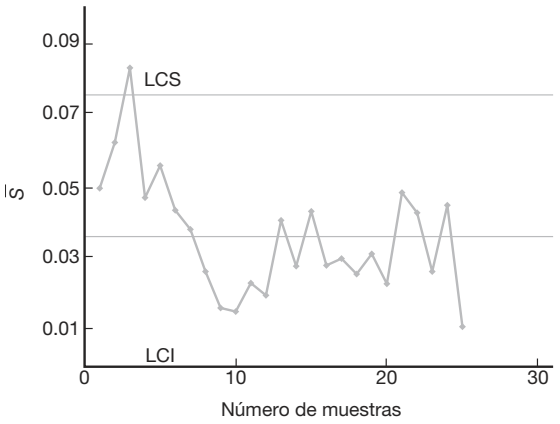


Figura 17.8: Gráfica $\bar{S}$ con límites de control establecidos con los datos del ejemplo 17.2.

Un estimador no sesgado de p es la **fracción de defectuosos** o la **proporción de defectuosos**, $\hat{p}$, donde

$$\hat{p} = \frac{\text{número de defectuosos en la muestra de tamaño } n}{n}.$$

Como en el caso de las gráficas de control de variables, las propiedades de distribución de p son importantes para la creación de la gráfica de control. Sabemos que

$$E(\hat{p}) = p, \qquad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

Aquí aplicamos los mismos principios 3σ que utilizamos para las gráficas de variables. Supongamos inicialmente que conocemos p. Entonces, la estructura de las gráficas de control implica utilizar límites 3σ con

$$\hat{\sigma} = \sqrt{\frac{p(1-p)}{n}}.$$

De esta manera, los límites son

$$\text{LCI} = p - 3\sqrt{\frac{p(1-p)}{n}}, \qquad \text{LCS} = p + 3\sqrt{\frac{p(1-p)}{n}},$$

con el proceso considerado bajo control cuando los valores $\hat{p}$ de la muestra caen dentro de los límites de control.

En general, por supuesto, no se conoce el valor de p y se debe estimar a partir de un conjunto base de muestras de forma muy similar al caso de μ y σ en las gráficas de variables. Suponga que hay m muestras preliminares de tamaño n. Para una muestra dada, cada una de las n observaciones se reporta como "defectuosa" o "no defectuosa". El estimador no sesgado evidente para p que se utiliza en la gráfica de control es

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_i,$$

donde $\hat{p}_i$ es la proporción de artículos defectuosos en la i-ésima muestra. Como resultado, los límites de control son

$$\text{LCI} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \qquad \text{línea central} = \bar{p}, \qquad \text{LCS} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

Ejemplo 17.3: Considere los datos que se presentan en la tabla 17.3 sobre el número de componentes electrónicos defectuosos en muestras de tamaño 50. Se tomaron 20 muestras con la finalidad de establecer valores preliminares para la gráfica de control. Las gráficas de control determinadas por este periodo preliminar tendrán una línea central $\hat{p} = 0.088$ y límites de control

$$\text{LCI} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{50}} = -0.0322 \quad \text{y} \quad \text{LCS} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{50}} = 0.2082.$$

Tabla 17.3: Datos para el ejemplo 17.3 que permiten establecer límites de control en gráficas p, con muestras de tamaño 50

Muestra	Número de componentes defectuosos	Fracción de defectuosos $\hat{p}_i$
1	8	0.16
2	6	0.12
3	5	0.10
4	7	0.14
5	2	0.04
6	5	0.10
7	3	0.06
8	8	0.16
9	4	0.08
10	4	0.08
11	3	0.06
12	1	0.02
13	5	0.10
14	4	0.08
15	4	0.08
16	2	0.04
17	3	0.06
18	5	0.10
19	6	0.12
20	3	0.06
		$\bar{p} = 0.088$

Evidentemente, con un valor calculado negativo, el LCI se ajusta a cero. A partir de los valores de los límites de control al parecer el proceso está bajo control durante este periodo preliminar.

Selección del tamaño de la muestra para la gráfica *p*

La elección del tamaño de la muestra para la gráfica p de atributos incluye los mismos tipos generales de consideraciones que los de la gráfica para variables. Se requiere un tamaño de la muestra tan grande como para tener una alta probabilidad de detectar una

condición fuera de control cuando, de hecho, ha ocurrido un cambio específico en p. No existe un *mejor método* para elegir el tamaño de la muestra. Sin embargo, Duncan (1986; véase la bibliografía) sugirió una aproximación razonable que consiste en elegir una n tal que haya 0.5 de probabilidades de detectar un cambio de una cantidad particular en p. La solución resultante para n es bastante simple. Suponga que se aplica la aproximación normal a la distribución binomial. Deseamos, siempre que la condición de p haya cambiado a, digamos, $p_1 > p_0$, que

$$P(\hat{p} \geq \text{LCS}) = P\left[Z \geq \frac{\text{LCS} - p_1}{\sqrt{p_1(1-p_1)/n}}\right] = 0.5.$$

Como $P(Z > 0) = 0.5$, se establece

$$\frac{\text{LCS} - p_1}{\sqrt{p_1(1-p_1)/n}} = 0.$$

Al sustituir,

$$p + 3\sqrt{\frac{p(1-p)}{n}} = \text{LCS},$$

tenemos

$$(p - p_1) + 3\sqrt{\frac{p(1-p)}{n}} = 0.$$

Ahora podemos calcular n, el tamaño de cada muestra:

$$n = \frac{9}{\Delta^2} p(1-p),$$

donde, desde luego, Δ es el "cambio" en el valor de p, y p es la probabilidad de un artículo defectuoso sobre la que se basan los límites de control. Sin embargo, si las gráficas de control se basan en límites $k\sigma$, entonces

$$n = \frac{k^2}{\Delta^2} p(1-p).$$

Ejemplo 17.4: Suponga que se diseña una gráfica de control de calidad de atributos con un valor de $p = 0.01$ para la probabilidad de tener bajo control un artículo defectuoso. ¿Cuál es el tamaño de la muestra por subgrupo que produce una probabilidad de 0.5 de que se detecte un cambio en el proceso para $p = p_1 = 0.05$? La gráfica p resultante incluirá límites 3σ.

***Solución*:** Aquí tenemos $\Delta = 0.04$. El tamaño adecuado de la muestra es

$$n = \frac{9}{(0.04)^2}(0.01)(0.99) = 55.69 \approx 56.$$

Gráficas de control para artículos defectuosos (uso del modelo de Poisson)

En el procedimiento anterior supusimos que el artículo bajo consideración es uno que está defectuoso, es decir, que no funciona**,** o uno que no tiene defecto, en cuyo caso el artículo funciona y, por lo tanto, es aceptable para el consumidor. En muchas situaciones este método del artículo "defectuoso o no" es demasiado simplista. Las unidades pueden contener defectos o no cumplir con las especificaciones, y aun así funcionar bastante bien para el consumidor. En realidad, en este caso sería importante ejercer control sobre el *número de defectos* o *número de artículos que no cumplen las especificaciones*. Este tipo de control de calidad tiene aplicación cuando las unidades no son simplistas ni grandes. Por ejemplo, el número de defectos puede ser muy útil como objeto de control cuando el artículo o unidad es, digamos, una computadora personal. Otro ejemplo es una unidad definida por 50 pies de tubería fabricada, donde el número de soldaduras defectuosas es el objeto del control de calidad; el número de defectos en 50 pies de alfombra fabricada o el número de "burbujas" en una hoja grande de vidrio fabricado.

A partir de lo aquí descrito queda claro que en este caso no es apropiada la distribución binomial. El número total de artículos que no cumplen las especificaciones en una unidad o el número promedio por unidad se podría usar como la medida para la gráfica de control. A menudo se supone que el número de artículos que no cumplen las especificaciones en una muestra tiene una distribución de Poisson. A este tipo de gráfica con frecuencia se le llama **gráfica *C***.

Suponga que el número de defectos X en una unidad de producto tiene una distribución de Poisson con parámetro λ. (Aquí $t = 1$ para el modelo de Poisson). Recuerde que para la distribución de Poisson,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

Aquí, la variable aleatoria X es el número de artículos que no cumplen las especificaciones. En el capítulo 5 vimos que tanto la media como la varianza de la variable aleatoria de Poisson son λ. Por consiguiente, si la gráfica de control de calidad se estructurara de acuerdo con los límites 3σ acostumbrados, si conociéramos λ tendríamos,

$$\text{LCS} = \lambda + 3\sqrt{\lambda}, \quad \text{línea central} = \lambda, \qquad \text{LCI} = \lambda - 3\sqrt{\lambda}.$$

Como de costumbre, λ a menudo debe provenir de un estimador de los datos. Un estimado no sesgado de λ es el número *promedio* de artículos que no cumplen las especificaciones por muestra. Este estimado se denota mediante $\hat{\lambda}$. Así, la gráfica de control tiene los límites

$$\text{LCS} = \hat{\lambda} + 3\sqrt{\hat{\lambda}}, \quad \text{línea central} = \hat{\lambda}, \qquad \text{LCI} = \hat{\lambda} - 3\sqrt{\hat{\lambda}}.$$

Ejemplo 17.5: La tabla 17.4 representa el número de defectos en 20 muestras sucesivas de rollos de hoja metálica, cada uno con 100 pies de largo. Para controlar el número de defectos en tales muestras se debe desarrollar una gráfica de control a partir de estos datos preliminares. El estimado del parámetro de Poisson λ es dado por $\hat{\lambda} = 5.95$. Como resultado, los límites de control sugeridos por estos datos preliminares son

$$\text{LCS} = \hat{\lambda} + 3\sqrt{\hat{\lambda}} = 13.2678 \quad \text{y} \quad \text{LCI} = \hat{\lambda} - 3\sqrt{\hat{\lambda}} = -1.3678,$$

donde LCI se iguala a cero.

Tabla 17.4: Datos para el ejemplo 17.5; el control implica el número de defectos en rollos de hojas metálicas

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	8	11	3
2	7	12	7
3	5	13	5
4	4	14	9
5	4	15	7
6	7	16	7
7	6	17	8
8	4	18	6
9	5	19	7
10	6	20	4
			Prom. 5.95

La figura 17.9 presenta una gráfica de los datos preliminares con los límites de control.

La tabla 17.5 incluye datos adicionales tomados del proceso de producción. Para cada muestra se inspeccionó la unidad en la que se basó la gráfica, a saber, 100 pies del metal. Se incluye la información de 20 muestras. La figura 17.10 muestra una gráfica de los datos adicionales de producción. Es evidente que el proceso está bajo control, o al menos lo estaba en el periodo en el que se tomaron los datos.

Tabla 17.5: Datos adicionales del proceso de producción del ejemplo 17.5

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	3	11	7
2	5	12	5
3	8	13	9
4	5	14	4
5	8	15	6
6	4	16	5
7	3	17	3
8	6	18	2
9	5	19	1
10	2	20	6

En el ejemplo 17.5 dejamos muy claro que la unidad de muestreo o de inspección son 100 pies de metal. En muchos casos en los que el artículo es específico, como en el caso de una computadora personal o el de un tipo específico de dispositivo electrónico, la unidad de inspección podría ser un *conjunto de artículos*. Por ejemplo, el analista decide utilizar 10 computadoras en cada subgrupo y de esta forma observar un conteo del número total de defectos encontrados. Por consiguiente, la muestra preliminar para construir la gráfica de control implica utilizar varias muestras, cada una de 10 computadoras. La elección del tamaño de la muestra puede depender de muchos factores. A menudo deseamos un tamaño de la muestra que asegure un LCI positivo.

El analista podría utilizar el número promedio de defectos por unidad de muestreo como la medida básica de la gráfica de control. Por ejemplo, para el caso de la compu-

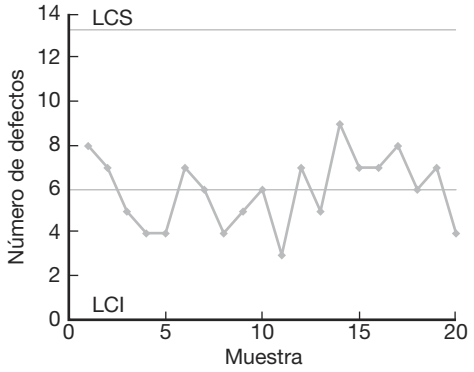


Figura 17.9: Datos preliminares representados en la gráfica de control para el ejemplo 17.5.

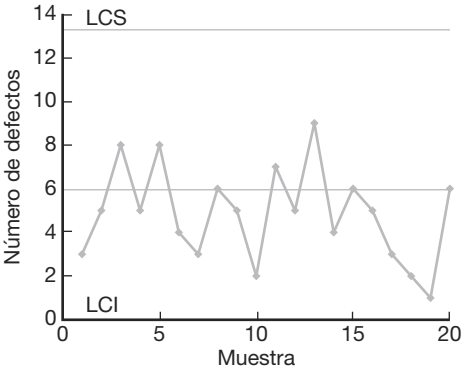


Figura 17.10: Datos adicionales de producción para el ejemplo 17.5.

tadora personal, sea la variable aleatoria el número total de defectos

$$U = \frac{\text{número total de defectos}}{n}$$

que se mide para cada muestra de, digamos, $n = 10$. Si suponemos que el número de defectos por unidad de muestreo es de Poisson con parámetro λ, podemos utilizar el método de las funciones generadoras de momento para demostrar que U es una variable aleatoria de Poisson (véase el ejercicio de repaso 17.1). De esta manera, la gráfica de control para esta situación se caracteriza por lo siguiente:

$$\text{LCS} = \bar{U} + 3\sqrt{\frac{\bar{U}}{n}}, \qquad \text{línea central} = \bar{U}, \qquad \text{LCI} = \bar{U} - 3\sqrt{\frac{\bar{U}}{n}}.$$

Aquí, desde luego, $\bar{U}$ es el promedio de los valores U en el conjunto de datos preliminares o base. El término $\bar{U}/n$ se deriva del resultado que

$$E(U) = \lambda, \qquad \text{Var}(U) = \frac{\lambda}{n},$$

y por ello $\bar{U}$ es un estimado no sesgado de $E(U) = \lambda$ y $\bar{U}/n$ es un estimado no sesgado de $\text{Var}(U) = \lambda/n$. Este tipo de gráfica de control a menudo se denomina **gráfica *U***.

En esta sección basamos toda la explicación de las gráficas de control en el modelo de probabilidad de Poisson. Este modelo se ha utilizado en combinación con el concepto 3σ. Como explicamos antes en este capítulo, el concepto de límites 3σ tiene sus raíces en la aproximación normal, aunque muchos usuarios consideran que el concepto funciona bien como herramienta pragmática incluso si la normalidad no es siquiera aproximadamente correcta. La dificultad, desde luego, radica en el hecho de que, en ausencia de normalidad, no es posible controlar la probabilidad de una especificación incorrecta de un estado fuera de control. En el caso del modelo de Poisson, cuando λ es pequeña la distribución es bastante asimétrica, una condición que puede producir resultados indeseables si se utiliza el método 3σ.

17.6 Gráficas de control de cusum

La desventaja de las gráficas de control similares a las de Shewhart, que se explicaron y ejemplificaron en las secciones anteriores, radica en su incapacidad para detectar pequeños cambios en la media. Un mecanismo de control de calidad que ha recibido mucha atención en la literatura estadística y que se ha utilizado extensamente en la industria es la **gráfica de suma acumulada (cusum)**. El método de la gráfica de suma acumulada es sencillo y, por lo tanto, atractivo. Para el lector debe ser evidente por qué es más sensible a pequeños cambios en la media. Considere una gráfica de control para la media con un nivel de referencia establecido en el valor W. Considere las observaciones particulares $X_1, X_2,..., X_r$. Las primeras cusum r son

$$\begin{aligned} S_1 &= X_1 - W \\ S_2 &= S_1 + (X_2 - W) \\ S_3 &= S_2 + (X_3 - W) \\ &\vdots \\ S_r &= S_{r-1} + (X_r - W). \end{aligned}$$

Es evidente que la cusum es simplemente la acumulación de las diferencias del nivel de referencia. Es decir,

$$S_k = \sum_{i=1}^{k} (X_i - W), \qquad k = 1, 2, \ldots.$$

La gráfica cusum es, entonces, una gráfica de S_k contra el tiempo.

Suponga que consideramos que el nivel de referencia W es un valor aceptable de la media μ. Salta a la vista que, si no hay cambio en μ, la gráfica cusum debería ser aproximadamente horizontal, con algunas fluctuaciones menores balanceadas alrededor de cero. Ahora, si sólo hay un cambio moderado en la media, debe resultar un cambio más o menos grande en la *pendiente* de la gráfica cusum, dado que cada nueva observación tiene la probabilidad de contribuir a un cambio y la medida que se grafica acumula esos cambios. Desde luego, la señal de que la media ha cambiado reside en la naturaleza de la pendiente de la gráfica cusum. El objetivo de la gráfica es detectar cambios que se alejan del nivel de referencia. Una pendiente diferente de cero (en cualquier dirección) representa un cambio a partir del nivel de referencia. Una pendiente positiva indica un aumento en la media por arriba del nivel de referencia, en tanto que una pendiente negativa señala una disminución.

Las gráficas cusum a menudo se diseñan con un *nivel de calidad aceptable* definido (NCA) y un *nivel de calidad rechazable* (NCR) preestablecido por el usuario. Ambos representan valores de la media. Se podría considerar que éstos desempeñan papeles similares a los de las medias nula y alternativa en la prueba de hipótesis. Considere una situación en la que el analista desea detectar un aumento en el valor de la media del proceso. Usaremos la notación μ_0 para NCA y μ_1 para NCR, y $\mu_1 > \mu_0$. El nivel de referencia se fija ahora en

$$W = \frac{\mu_0 + \mu_1}{2}.$$

Los valores de S_r ($r = 1, 2,...$) tendrán una pendiente negativa si la media del proceso está en μ_0 y una pendiente positiva si la media del proceso está en μ_1.

Regla de decisión para las gráficas cusum

Como antes se expuso, la pendiente de la gráfica cusum proporciona la señal de acción para el analista de control de calidad. La regla de decisión exige tomar medidas si, en el r-ésimo periodo de muestreo,

$$d_r > h,$$

donde h es un valor preestablecido que se denomina **longitud del intervalo de decisión** y

$$d_r = S_r - \min_{1 \le i \le r-1} S_i.$$

En otras palabras, se toman medidas si los datos revelan que el valor de la cusum real excede en una cantidad específica al valor previo de la cusum más pequeño.

Una modificación en la mecánica que se describió antes facilita el uso del método. Describimos un procedimiento que grafica las cusum y calcula las diferencias. Una modificación simple implica graficar las diferencias de manera directa y permitir la verificación contra el intervalo de decisión. La expresión general para d_r es muy sencilla. Para el procedimiento de cusum, con el que se detectan aumentos en la media,

$$d_r = \text{máx}[0, d_{r-1} + (X_r - W)].$$

La elección del valor de h es, por supuesto, muy importante. En este libro no se proporcionan los detalles que aparecen en la literatura que trata de esta elección. Para una exposición más completa se remite al lector a Ewan y Kemp, 1960, y a Montgomery, 2000b (véase la bibliografía). Una consideración importante es la **longitud esperada de la corrida**. De manera ideal, la longitud esperada de la corrida es bastante grande bajo $\mu = \mu_0$ y muy pequeña cuando $\mu = \mu_1$.

Ejercicios de repaso

17.1 Considere $X_1, X_2, \ldots, X_n$, como variables aleatorias de Poisson independientes con parámetros $\mu_1, \mu_2, \ldots, \mu_n$. Utilice las propiedades de las funciones generadoras de momento para demostrar que la variable aleatoria $\sum_{i=1}^{n} X_i$ es una variable aleatoria de Poisson con media $\sum_{i=1}^{n} \mu_i$ y varianza $\sum_{i=1}^{n} \mu_i$.

17.2 Considere los siguientes datos tomados en subgrupos de tamaño 5. Los datos contienen 20 promedios y rangos del diámetro (en milímetros) de una parte importante de un motor. Elabore gráficas $\bar{X}$ y R. ¿Parecería que el proceso está bajo control?

Muestra	$\bar{X}$	R
1	2.3972	0.0052
2	2.4191	0.0117
3	2.4215	0.0062
4	2.3917	0.0089
5	2.4151	0.0095
6	2.4027	0.0101
7	2.3921	0.0091
8	2.4171	0.0059
9	2.3951	0.0068
10	2.4215	0.0048
11	2.3887	0.0082
12	2.4107	0.0032
13	2.4009	0.0077
14	2.3992	0.0107
15	2.3889	0.0025
16	2.4107	0.0138
17	2.4109	0.0037
18	2.3944	0.0052
19	2.3951	0.0038
20	2.4015	0.0017

17.3 En el ejercicio de repaso 17.2 suponga que el comprador fija especificaciones para la parte. Las especificaciones exigen que el diámetro caiga en el rango cubierto por 2.40000 ± 0.0100 mm. ¿Qué proporción de unidades producidas por este proceso no cumplirán con las especificaciones?

17.4 Para la situación del ejercicio de repaso 17.2 proporcione estimados numéricos de la media y de la desviación estándar del diámetro para la parte que se fabrica en el proceso.

17.5 Considere los datos de la tabla 17.1. Suponga que se toman muestras adicionales de tamaño 5 y que se registra la resistencia a la tensión. El muestreo produce los siguientes resultados (en libras por pulgada cuadrada).

Muestra	$\bar{X}$	R
1	1511	22
2	1508	14
3	1522	11
4	1488	18
5	1519	6
6	1524	11
7	1519	8
8	1504	7
9	1500	8
10	1519	14

a) Grafique los datos, utilice las gráficas $\bar{X}$ y R para los datos preliminares de la tabla 17.1.
b) ¿Parecería que el proceso está bajo control? Si no es así, explique por qué.

17.6 Considere un proceso bajo control con media $\mu = 25$ y $\sigma = 1.0$. Suponga que se usan subgrupos de tamaño 5 con límites de control $\mu \pm 3\sigma/\sqrt{n}$ y línea central en μ. Suponga que ocurre un cambio en la media, y que la nueva media es $\mu = 26.5$.
a) ¿Cuál es el número promedio de muestras requeridas (después del cambio) para detectar la situación fuera de control?
b) ¿Cuál es la desviación estándar del número de corridas requeridas?

17.7 Considere la situación del ejemplo 17.2. Se toman los siguientes datos de muestras adicionales de tamaño 5. Grafique los valores $\bar{X}$ y S sobre las gráficas $\bar{X}$ y S que se dibujaron con los datos en la muestra preliminar. ¿Parecería que el proceso está bajo control? Explique su respuesta.

Muestra	$\bar{X}$	S_i
1	62.280	0.062
2	62.319	0.049
3	62.297	0.077
4	62.318	0.042
5	62.315	0.038
6	62.389	0.052
7	62.401	0.059
8	62.315	0.042
9	62.298	0.036
10	62.337	0.068

17.8 Cada hora se toman muestras de tamaño 50 de un proceso que produce cierto tipo de artículo que se considera que está defectuoso o que no tiene defecto. Se toman 20 muestras.

a) Construya una gráfica de control para controlar la proporción de artículos defectuosos.
b) ¿Parecería que el proceso está bajo control? Explique su respuesta.

Muestra	Número de artículos defectuosos	Muestra	Número de artículos defectuosos
1	4	11	2
2	3	12	4
3	5	13	1
4	3	14	2
5	2	15	3
6	2	16	1
7	2	17	1
8	1	18	2
9	4	19	3
10	3	20	1

17.9 Para la situación del ejercicio de repaso 17.8 suponga que se reúnen los siguientes datos adicionales:

Muestra	Número de artículos defectuosos
1	3
2	4
3	2
4	2
5	3
6	1
7	3
8	5
9	7
10	7

¿Parecería que el proceso está bajo control? Explique su respuesta.

17.10 Se aplica un programa de control de calidad para un proceso, donde se fabrican grandes placas de acero, con un interés especial por los defectos superficiales. El objetivo es establecer una gráfica de control de calidad para el número de defectos por placa. Los datos se presentan a continuación. Elabore la gráfica de control apropiada utilizando esta información. ¿Parecería que el proceso está bajo control?

Muestra	Número de defectos	Muestra	Número de defectos
1	4	11	1
2	2	12	2
3	1	13	2
4	3	14	3
5	0	15	1
6	4	16	4
7	5	17	3
8	3	18	2
9	2	19	1
10	2	20	3

Capítulo 18

Estadística bayesiana

18.1 Conceptos bayesianos

Los métodos clásicos de estimación que hemos estudiado hasta ahora se basan sólo en la información que brinda la muestra aleatoria. Estos métodos en esencia interpretan probabilidades como frecuencias relativas. Por ejemplo, para obtener un intervalo de confianza de 95% para μ, interpretamos la aseveración

$$P(-1.96 < Z < 1.96) = 0.95$$

para afirmar que, en experimentos repetidos, Z caerá 95% de las veces entre -1.96 y 1.96. Dado que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

para una muestra normal con varianza conocida, el enunciado de probabilidad aquí significa que 95% de los intervalos aleatorios $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ contienen la media μ verdadera. Otro enfoque de los métodos estadísticos de estimación se denomina **metodología bayesiana**. La idea principal del método proviene de la regla de Bayes, que examinamos en la sección 2.7. La diferencia fundamental entre el enfoque bayesiano y el clásico o frecuente es que en los conceptos bayesianos los parámetros se consideran variables aleatorias.

Probabilidad subjetiva

La probabilidad subjetiva es el fundamento de los conceptos bayesianos. En el capítulo 2 analizamos dos acercamientos posibles a la probabilidad, es decir, el método de la frecuencia relativa y el método de la indiferencia. El primero determina una probabilidad como una consecuencia de experimentos repetidos. Por ejemplo, para decidir el porcentaje de tiros libres de un jugador de basquetbol, podemos registrar el número de tiros que hace y el número total de intentos que tal jugador ha hecho hasta el momento. La probabilidad de que este jugador acierte un tiro libre se puede calcular como el cociente de estos dos números. Por otro lado, si no sabemos acerca de cualquier sesgo en un dado, la probabilidad de que aparezca un 3 en el siguiente lanzamiento será de 1/6. Dicho enfoque en la interpretación de la probabilidad se basa en la regla de la indiferencia.

Sin embargo, en muchas situaciones no es posible aplicar las interpretaciones de probabilidad anteriores. Por ejemplo, considere las siguientes preguntas: "¿Qué probabilidad hay de que llueva mañana?" "¿Qué tan probable es que el precio de estas acciones aumente a fin de mes?" y "¿Cuál es la probabilidad de que dos empresas se fusionen?". Estas preguntas difícilmente se podrían interpretar mediante los enfoques anteriores, y las respuestas podrían ser diferentes para distintas personas. No obstante, este tipo de preguntas se plantean constantemente en la vida diaria y el enfoque utilizado para explicar esas probabilidades se llama *probabilidad subjetiva*, ya que refleja opiniones subjetivas.

Perspectiva condicional

Recuerde que en los capítulos 9 a 17 todas las inferencias estadísticas se basaban en el hecho de que los parámetros se desconocen pero son cantidades fijas, excepto los revisados en la sección 9.14, en donde los parámetros se trataron como variables y los estimados de máxima verosimilitud (EMV) se calcularon con base en la muestra de datos observados. En la estadística bayesiana los parámetros no sólo se manejan como variables, como en los cálculos de EMV, sino que también se manejan como aleatorios.

Puesto que los datos observados son los únicos resultados experimentales para el profesionista, la inferencia estadística se basa en los datos reales observados a partir de un experimento dado. A esta visión se le llama *perspectiva condicional*. Más aún, en los conceptos bayesianos, dado que los parámetros se manejan como aleatorios, es factible especificar una distribución de probabilidad, por lo general utilizando la *probabilidad subjetiva* para el parámetro. Este tipo de distribución se denomina *distribución previa* y comúnmente refleja la creencia previa del experimentador acerca del parámetro. En la perspectiva bayesiana, una vez que se realiza un experimento y se observan los datos, todo el conocimiento acerca de un parámetro está contenido en los datos reales observados, así como en la información previa.

Aplicaciones bayesianas

Aunque la regla de Bayes se atribuye a Thomas Bayes, las aplicaciones bayesianas fueron utilizadas por primera vez por el científico francés Pierre Simon Laplace, quien publicó un artículo sobre el uso de la inferencia bayesiana en las proporciones binomiales desconocidas (para revisar la distribución binomial véase la sección 5.2).

A partir de la introducción del paquete para el cálculo de la cadena Markov de Monte Carlo (MCMC) para el análisis bayesiano a principios de la década de 1990, los métodos bayesianos se han vuelto cada vez más populares para los modelos estadísticos y el análisis de datos. Al mismo tiempo, la metodología que utiliza conceptos bayesianos ha avanzado mucho y se aplica en campos como la bioinformática, la biología, los negocios, la ingeniería, las ciencias ambientales y la ecología, así como en la ciencia de la vida y la salud, entre otros.

18.2 Inferencias bayesianas

Considere el problema de calcular un estimado puntual del parámetro θ para la población con distribución $f(x|\theta)$, dado θ. Denote con $\pi(\theta)$ la distribución previa de q. Suponga que se observa una muestra aleatoria de tamaño n denotada con $\mathbf{x} = (x_1, x_2, ..., x_n)$.

Definición 18.1: La distribución de θ, dado $\mathbf{x}$, que se denomina distribución posterior, es dada por

$$\pi(\theta\,|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{g(\mathbf{x})},$$

donde $g(\mathbf{x})$ es la distribución marginal de $\mathbf{x}$.

La distribución marginal de $\mathbf{x}$ en la definición anterior se puede calcular usando la siguiente fórmula:

$$g(\mathbf{x}) = \begin{cases} \sum\limits_{\theta} f(\mathbf{x}|\theta)\pi(\theta), & \theta \text{ es discreta,} \\ \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)\pi(\theta)\,d\theta, & \theta \text{ es continua.} \end{cases}$$

Ejemplo 18.1: Suponga que la distribución previa para la proporción de artículos defectuosos que produce una máquina es

p	0.1	0.2
$\pi(p)$	0.6	0.4

Denote con x el número de artículos defectuosos en una muestra aleatoria de tamaño 2. Calcule la distribución de probabilidad posterior de p, dado que se observa x.

***Solución*:** La variable aleatoria X sigue una distribución binomial

$$f(x|p) = b(x;2,p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

La distribución marginal de x se puede calcular como

$$\begin{aligned} g(x) &= f(x|0.1)\pi(0.1) + f(x|0.2)\pi(0.2) \\ &= \binom{2}{x}[(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)]. \end{aligned}$$

Por lo tanto, para $x = 0, 1, 2$ obtenemos las siguientes probabilidades marginales

x	0	1	2
$g(x)$	0.742	0.236	0.022

La probabilidad posterior de $p = 0.1$, dado x, es

$$\pi(0.1|x) = \frac{f(x|0.1)\pi(0.1)}{g(x)} = \frac{(0.1)^x(0.9)^{2-x}(0.6)}{(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)},$$

y $\pi(0.2|x) = 1 - \pi(0.1|x)$.

Suponga que se observa $x = 0$.

$$\pi(0.1|0) = \frac{f(0 \mid 0.1)\,\pi(0.1)}{g(0)} = \frac{(0.1)^0(0.9)^{2-0}(0.6)}{0.742} = 0.6550,$$

y $\pi(0.2|0) = 0.3450$. Si se observa $x = 1$, $\pi(0.1|1) = 0.4576$ y $\pi(0.2|1) = 0.5424$. Por último, $\pi(0.1|2) = 0.2727$ y $\pi(0.2|2) = 0.7273$. ◢

La distribución previa del ejemplo 18.1 es discreta, aunque el rango natural de p va de 0 a 1. Considere el siguiente ejemplo, en el cual tenemos una distribución previa que abarca el espacio completo de p.

Ejemplo 18.2: Suponga que la distribución previa de p es uniforme, es decir, $\pi(p) = 1$, para $0 < p < 1$. Use la misma variable aleatoria X que en el ejemplo 18.1 para calcular la distribución posterior de p.

***Solución*:** Como en el ejemplo 18.1, tenemos

$$f(x|p) = b(x; 2, p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

La distribución marginal de x se puede calcular como

$$g(x) = \int_0^1 f(x|p)\pi(p)\, dp = \binom{2}{x} \int_0^1 p^x (1-p)^{2-x}\, dp.$$

La integral anterior se puede evaluar en cada x directamente como $g(0) = 1/3$, $g(1) = 1/3$ y $g(2) = 1/3$. Por lo tanto, la distribución posterior de p, dado x, es

$$\pi(p|x) = \frac{\binom{2}{x} p^x (1-p)^{2-x}}{1/3} = 3\binom{2}{x} p^x (1-p)^{2-x}, \quad 0 < p < 1.$$

La distribución posterior anterior es en realidad una distribución beta (véase la sección 6.8) con parámetros $\alpha = x + 1$ y $\beta = 3 - x$. Por lo tanto, si se observa $x = 0$, la distribución posterior de p es una distribución beta con parámetros (1, 3). La media posterior es $\mu = \frac{1}{1+3} = \frac{1}{4}$ y la varianza posterior es $\sigma^2 = \frac{(1)(3)}{(1+3)^2\,(1+3+1)} = \frac{3}{80}$. ∎

Si utilizamos la distribución posterior, podemos estimar directamente el (los) parámetro(s) en una población. Al calcular las distribuciones posteriores es muy útil estar familiarizado con las distribuciones que se estudiaron en los capítulos 5 y 6. Observe que en la definición 18.1 la *variable* en la distribución posterior es θ, en tanto se proporcione $\mathbf{x}$. Por consiguiente, podemos tratar a $g(\mathbf{x})$ como una constante cuando calculamos la distribución posterior de θ. Entonces, la distribución posterior se puede expresar como

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta),$$

donde el símbolo "$\propto$" significa *proporcional a*. En el cálculo que se hizo de la distribución posterior podríamos dejar los factores que no dependen de θ fuera de la constante de normalización, esto es, la densidad marginal $g(\mathbf{x})$.

Ejemplo 18.3 Suponga que las variables aleatorias X_1,..., X_n son independientes y provienen de una distribución de Poisson con media λ. Suponga que la distribución previa de λ es exponencial con media 1. Calcule la distribución posterior de λ cuando $\bar{x} = 3$ con $n = 10$.

***Solución*:** La función de densidad de $\mathbf{X} = (X_1,..., X_n)$ es

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!},$$

y la distribución previa es

$$\pi(\theta) = e^{-\lambda}, \text{ para } \lambda > 0.$$

En consecuencia, utilizando la definición 18.1 se obtiene la siguiente distribución posterior de λ

$$\pi(\lambda|\mathbf{x}) \propto f(\mathbf{x}|\lambda)\pi(\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}e^{-\lambda} \propto e^{-(n+1)\lambda}\lambda^{\sum_{i=1}^{n} x_i}.$$

Si nos remitimos a la distribución gamma en la sección 6.6, concluimos que la distribución posterior de λ sigue una distribución gamma con parámetros $1 + \sum_{i=1}^{n} x_i$ y $\frac{1}{n+1}$. Por lo tanto, tenemos la media y la varianza posterior de λ como $\frac{\sum_{i=1}^{n} x_i + 1}{n+1}$ y $\frac{\sum_{i=1}^{n} x_i + 1}{(n+1)^2}$. Así, cuando $\bar{x} = 3$ con $n = 10$, tenemos $\sum_{i=1}^{10} x_i = 30$. Por lo tanto, la distribución posterior de λ es una distribución gamma con parámetros 31 y 1/11. ∎

A partir del ejemplo 18.3 observamos que en ocasiones es muy conveniente usar la técnica "proporcional a" para calcular la distribución posterior, especialmente cuando el resultado se puede formar para una distribución de uso común como las que se describen en los capítulos 5 y 6.

Estimación puntual mediante la distribución posterior

Una vez que hemos derivado la distribución posterior, fácilmente podemos usar el resumen de la distribución posterior para hacer inferencias sobre los parámetros de la población. Por ejemplo, la media, la mediana y la moda posteriores son útiles para estimar el parámetro.

Ejemplo 18.4: Suponga que en el ejemplo 18.2 se observa $x = 1$. Determine la media y la moda posteriores.

***Solución*:** Cuando $x = 1$, la distribución posterior de p se puede expresar como

$$\pi(p|1) = 6p(1-p), \quad \text{para} \quad 0 < p < 1.$$

Para calcular la media de esta distribución necesitamos encontrar

$$\int_0^1 6p^2(1-p)\,dp = 6\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{1}{2}.$$

Para determinar la moda posterior se requiere obtener un valor p tal que se maximice la distribución posterior. Si tomamos la derivada de $\pi(p)$ respecto a p, obtenemos $6 - 12p$. Al despejar p en $6 - 12p = 0$, obtenemos $p = ½$. La segunda derivada es -12, la cual implica que la moda posterior se logra en $p = ½$. ∎

Los métodos bayesianos de estimación respecto a la media μ de una población normal se basan en el siguiente ejemplo.

Ejemplo 18.5: Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n tomada de una población normal con varianza conocida σ^2, y la distribución previa de la media poblacional es una distribución normal con media conocida μ_0 y varianza conocida σ_0^2, demuestre que la distribución

posterior de la media poblacional es también una distribución normal con media μ^* y desviación estándar σ^*, donde

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 \qquad \text{y} \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

***Solución*:** La función de densidad de la muestra es

$$f(x_1, x_2, \ldots, x_n \mid \mu) = \frac{1}{(2\pi)^{n/2}\,\sigma^n}\,\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right],$$

para $-\infty < x_i < \infty$ e $i = 1, 2,...,n$, y la previa es

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\,\exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right], \qquad -\infty < \mu < \infty.$$

Entonces la distribución posterior de μ es

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{n(\bar{x} - \mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\right]\right\},\end{aligned}$$

debido a

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

de la sección 8.5. Al completar los cuadrados para μ se obtiene la distribución posterior

$$\pi(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu^*}{\sigma^*}\right)^2\right],$$

donde

$$\mu^* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

Ésta es una distribución normal con media μ^* y desviación estándar σ^*. ∎

El teorema del límite central nos permite utilizar el ejemplo 18.5 también cuando seleccionamos muestras aleatorias suficientemente grandes ($n \geq 30$ para muchos casos de experimentación en ingeniería), a partir de poblaciones no normales (la distribución no dista mucho de ser simétrica), y cuando la distribución previa de la media es aproximadamente normal.

Resulta pertinente hacer algunos comentarios acerca del ejemplo 18.5. La media posterior μ^* también se puede escribir como

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0,$$

que es el promedio ponderado de la media muestral $\bar{x}$ y la media previa μ_0. Como ambos coeficientes están entre 0 y 1 y se suman a 1, la media posterior μ^* siempre se encuentra

entre $\bar{x}$ y μ_0. Esto significa que tanto $\bar{x}$ como μ_0 influyen en la estimación posterior de μ. Además, la ponderación de $\bar{x}$ depende de la varianza previa, así como de la varianza de la media muestral. Para un problema con una muestra grande ($n \to \infty$), la media posterior $\mu^* \to \bar{x}$. Esto significa que la media previa no desempeña ningún papel en la estimación de la media poblacional μ utilizando la distribución posterior. Esto es muy razonable, puesto que indica que cuando una cantidad de datos es sustancial, la información a partir de los datos dominará la información de μ proporcionada por la previa. Por otro lado, cuando la varianza previa es grande ($\sigma_0^2 \to \infty$), la media posterior μ^* también va hacia $\bar{x}$. Observe que para una distribución normal, cuanto mayor es la varianza, más plana será la función de densidad. El carácter plano de la distribución normal en este caso significa que casi no hay información previa subjetiva disponible del parámetro μ antes de reunir los datos. Por lo tanto, es razonable que la estimación posterior μ^* sólo dependa del valor de los datos de $\bar{x}$.

Ahora considere la desviación estándar posterior σ^*. Este valor también se escribe como

$$\sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2/n}{\sigma_0^2 + \sigma^2/n}}.$$

Es evidente que el valor σ^* es menor que σ_0 y que $\sigma/\sqrt{n}$, la desviación estándar previa y la desviación estándar de $\bar{x}$, respectivamente. Esto sugiere que la estimación posterior es más precisa que la previa y que los datos muestrales. En consecuencia, la incorporación tanto de los datos como de la información previa produce una mejor información posterior que si se utiliza cualquiera de los datos o la información previa por sí solos. Esto es un fenómeno común en la inferencia bayesiana. Además, para calcular μ^* y σ^* mediante las fórmulas del ejemplo 18.5 suponemos que se conoce σ^2. Como por lo general éste no es el caso, deberemos reemplazar σ^2 por la varianza de la muestra s^2 siempre que $n \geq 30$.

Estimación del intervalo bayesiano

De manera similar al intervalo de confianza clásico, en el análisis bayesiano podemos calcular un intervalo bayesiano del $100(1 - \alpha)\%$ empleando la distribución posterior.

Definición 18.2: El intervalo $a < \theta < b$ se denomina **intervalo de Bayes** del $100(1 - \alpha)\%$ para θ si

$$\int_{-\infty}^{a} \pi(\theta|x)\ d\theta = \int_{b}^{\infty} \pi(\theta|x)\ d\theta = \frac{\alpha}{2}.$$

Recuerde que, de acuerdo con el enfoque frecuentista, la probabilidad de un intervalo de confianza, digamos de 95%, se interpreta como una probabilidad de cobertura, esto significa que, si un experimento se repite una y otra vez (con considerables datos no observados), la probabilidad de que los intervalos calculados, de acuerdo con la regla, cubran el parámetro verdadero es de 95%. Sin embargo, en la interpretación del intervalo bayesiano, digamos para un intervalo de 95%, podemos decir que la probabilidad de que el parámetro desconocido caiga dentro del intervalo calculado (que sólo depende de los datos observados) es de 95%.

Ejemplo 18.6: Suponga que $X \sim b(x; n, p)$ con $n = 2$ conocida, y la distribución previa de p es uniforme $\pi(p) = 1$ para $0 < p < 1$. Calcule el intervalo de Bayes de 95% para p.

Solución: Como en el ejemplo 18.2, cuando $x = 0$ la distribución posterior es una distribución beta con parámetros 1 y 3, es decir, $\pi(p|0) = 3(1-p)^2$, para $0 < p < 1$. Por consiguiente, necesitamos despejar a y b utilizando la definición 18.2, lo que produce lo siguiente:

$$0.025 = \int_0^a 3(1-p)^2\, dp = 1 - (1-a)^3$$

y

$$0.025 = \int_b^1 3(1-p)^2\, dp = (1-b)^3.$$

Las soluciones a las ecuaciones anteriores dan como resultado $a = 0.0084$ y $b = 0.7076$. Por lo tanto, la probabilidad de que p caiga dentro de (0.0084, 0.7076) es de 95%. ∎

Para la población normal y el caso previo normal descrito en el ejemplo 18.5, la media posterior μ^* es el estimado de Bayes de la media poblacional μ, y se puede construir un **intervalo bayesiano** para μ de $100(1-\alpha)\%$ calculando el intervalo

$$\mu^* - z_{\alpha/2}\sigma^* < \mu < \mu^* + z_{\alpha/2}\sigma^*,$$

que se centra en la media posterior y contiene $100(1-\alpha)\%$ de la probabilidad posterior.

Ejemplo 18.7: Una empresa de equipo eléctrico fabrica bombillas con una duración distribuida de forma aproximadamente normal y una desviación estándar de 100 horas. La experiencia previa nos conduce a creer que μ es un valor de una variable aleatoria normal con una media $\mu_0 = 800$ horas y una desviación estándar $\sigma_0 = 10$ horas. Si una muestra aleatoria de 25 bombillas tiene una duración promedio de 780 horas, calcule un intervalo bayesiano de 95% para μ.

***Solución*:** De acuerdo con el ejemplo 18.5, la distribución posterior de la media también es una distribución normal con media

$$\mu^* = \frac{(25)(780)(10)^2 + (800)(100)^2}{(25)(10)^2 + (100)^2} = 796$$

y desviación estándar

$$\sigma^* = \sqrt{\frac{(10)^2(100)^2}{(25)(10)^2 + (100)^2}} = \sqrt{80}.$$

El intervalo bayesiano de 95% para μ es dado entonces por

$$796 - 1.96\sqrt{80} < \mu < 796 + 1.96\sqrt{80},$$

o

$$778.5 < \mu < 813.5.$$

En consecuencia, estamos 95% seguros de que μ estará entre 778.5 y 813.5.

Por otro lado, si desconocemos la información previa acerca de μ, procedemos como en la sección 9.4 para construir el intervalo de confianza clásico de 95%.

$$780 - (1.96)\left(\frac{100}{\sqrt{25}}\right) < \mu < 780 + (1.96)\left(\frac{100}{\sqrt{25}}\right),$$

o $740.8 < \mu < 819.2$, el cual se ve que es más amplio que el intervalo bayesiano correspondiente. ∎

18.3 Estimados bayesianos mediante la teoría de decisión

Con la metodología bayesiana se puede obtener la distribución posterior del parámetro. Los estimados bayesianos también se pueden derivar usando la distribución posterior y una función de pérdida cuando se incurre en una pérdida. Una función de pérdida es aquella que describe el costo de una decisión asociada con un suceso de interés. Aquí sólo se citan unas cuantas funciones de pérdida de uso común y sus estimados de Bayes asociados.

Pérdida del cuadrado del error

Definición 18.3: La **función de pérdida del cuadrado del error** es

$$L(\theta, a) = (\theta - a)^2,$$

donde θ es el parámetro (o estado natural) y a una acción (o estimado).

Un estimado de Bayes minimiza la pérdida posterior esperada dada en los datos muestrales observados.

Teorema 18.1: La media de la distribución posterior $\pi(\theta|x)$, denotada con θ^*, es el **estimado de Bayes de** θ bajo la función de pérdida del cuadrado del error.

Ejemplo 18.8: Calcule el estimado de Bayes de p para todos los valores de x en el ejemplo 18.1 cuando se utiliza la función de pérdida del cuadrado del error.

***Solución*:** Cuando $x = 0$, $p^* = (0.1)(0.6550) + (0.2)(0.3450) = 0.1345$.
Cuando $x = 1$, $p^* = (0.1)(0.4576) + (0.2)(0.5424) = 0.1542$.
Cuando $x = 2$, $p^* = (0.1)(0.2727) + (0.2)(0.7273) = 0.1727$.

Observe que el estimado clásico de p es $\hat{p} = x/n = 0$, ½ y 1, respectivamente, para los valores de x en 0, 1 y 2. Estos estimados clásicos son muy diferentes de los estimados de Bayes correspondientes. ■

Ejemplo 18.9: Repita el ejemplo 18.8 en la situación del ejemplo 18.2.

***Solución*:** Puesto que la distribución posterior de p es una distribución $B(x + 1, 3 - x)$ (véase la sección 6.8 en la página 201), el estimado de Bayes de p es

$$p^* = E^{\pi(p|x)}(p) = 3\binom{2}{x}\int_0^1 p^{x+1}(1-p)^{2-x}\,dp,$$

que produce $p^* = ¼$ para $x = 0$, $p^* = ½$ para $x = 1$, y $p^* = ¾$ para $x = 2$, respectivamente. Advierta que cuando se observa $x = 1$, el estimado de Bayes y el estimado clásico de $\hat{p}$ son equivalentes. ■

Para la situación normal que se describe en el ejemplo 18.5 el estimado de Bayes de μ bajo la pérdida del cuadrado del error será la media posterior μ^*.

Ejemplo 18.10: Suponga que la distribución muestral de una variable aleatoria X es de Poisson con parámetro λ. Suponga que la distribución previa de λ sigue una distribución gamma con

parámetros (α, β). Calcule el estimado de Bayes de λ bajo la función de pérdida del cuadrado del error.

***Solución*:** Si utilizamos el ejemplo 18.3, concluimos que la distribución posterior de λ sigue una distribución gamma con parámetros $(x + \alpha, (1 + 1/\beta)^{-1})$. Por medio del teorema 6.4 obtenemos la media posterior

$$\hat{\lambda} = \frac{x + \alpha}{1 + 1/\beta}.$$

Como la media posterior es el estimado de Bayes bajo la pérdida del cuadrado del error, $\hat{\lambda}$ es nuestro estimado de Bayes. ∎

Pérdida del error absoluto

La pérdida del cuadrado del error descrita antes es similar al concepto de los mínimos cuadrados que se analizó en relación con la regresión en los capítulos 11 y 12. En esta sección presentamos otra función de pérdida como sigue.

Definición 18.4: La **función de pérdida del error absoluto** se define como

$$L(\theta, a) = |\theta - a|,$$

donde θ es el parámetro y a una acción.

Teorema 18.2: La mediana de la distribución posterior $\pi(\theta|x)$, denotada con θ^*, es el **estimado de Bayes de** θ bajo la función de pérdida del error absoluto.

Ejemplo 18.11: Bajo la pérdida del error absoluto calcule el estimador de Bayes para el ejemplo 18.9 cuando se observa $x = 1$.

***Solución*:** Nuevamente, la distribución posterior de p es $B(x + 1, 3 - x)$. Cuando $x = 1$ se trata de una distribución beta con densidad $\pi(p \mid x = 1) = 6x(1 - x)$ para $0 < x < 1$ y 0 en otro caso. La mediana de esta distribución es un valor de p^* tal que

$$\frac{1}{2} = \int_0^{p^*} 6p(1 - p)\, dp = 3p^{*2} - 2p^{*3},$$

que produce la respuesta $p^* = \frac{1}{2}$. Por lo tanto, el estimado de Bayes en este caso es 0.5.

Ejercicios

18.1 Estime la proporción de artículos defectuosos que produce la máquina del ejemplo 18.1 si la muestra aleatoria de tamaño 2 produce dos artículos defectuosos.

18.2 Supongamos que la distribución previa para la proporción p de bebidas de una máquina despachadora que se derraman al servirse es

p	0.05	0.10	0.15
$\pi(p)$	0.3	0.5	0.2

Si dos de las siguientes 9 bebidas de esta máquina se derraman, calcule

a) la distribución posterior para la proporción p;
b) el estimado de Bayes de p.

18.3 Repita el ejercicio 18.2 cuando una de las siguientes 4 bebidas se derrama y la distribución uniforme previa es

$$\pi(p) = 10, \quad 0.05 < p < 0.15.$$

18.4 Las llamadas de servicio llegan a un centro de mantenimiento de acuerdo con un proceso de Poisson con λ llamadas por minuto. Un conjunto de datos de 20 periodos de un minuto producen un promedio de 1.8 llamadas. Si la distribución previa de λ sigue una distribución exponencial con media 2, determine la distribución posterior de λ.

18.5 Un estudio previo indica que el porcentaje de fumadores empedernidos, p, que tienen cáncer de pulmón sigue una distribución beta (véase la sección 6.8) con media de 70% y desviación estándar de 10%. Suponga que un nuevo conjunto de datos recolectado indica que 81 de 120 fumadores empedernidos tiene cáncer de pulmón.

a) Determine la distribución posterior del porcentaje de fumadores empedernidos que tienen cáncer de pulmón combinando los nuevos datos y la información previa.

b) ¿Cuál es la probabilidad posterior de que p sea mayor que 50%?

18.6 El constructor de un nuevo complejo de condominios afirma que 3 de 5 compradores preferirá un departamento de dos recámaras, mientras que su banquero afirma que sería más correcto decir que 7 de 10 compradores preferirán uno de dos recámaras. En las predicciones previas de este tipo el banquero ha sido dos veces más confiable que el constructor. Si 12 de los siguientes 15 condominios que se venden en este complejo son de dos recámaras, calcule

a) las probabilidades posteriores que se asocian con las afirmaciones del constructor y del banquero;

b) un estimado puntual de la proporción de compradores que prefieren un condominio de dos recámaras.

18.7 El tiempo en que se consume la primera etapa de un cohete es una variable aleatoria normal con una desviación estándar de 0.8 minutos. Suponga una distribución previa normal para μ con una media de ocho minutos y una desviación estándar de 0.2 minutos. Si se lanzan 10 de estos cohetes y la primera etapa tiene un tiempo de consumo promedio de 9 minutos, calcule un intervalo bayesiano de 95% para μ.

18.8 La utilidad diaria de una máquina despachadora de jugos, ubicada en un edificio de oficinas, es un valor de una variable aleatoria normal, con media μ y varianza σ^2 desconocidas. Desde luego, la media variará un poco de un edificio a otro, y el distribuidor considera que estas utilidades promedio diarias se pueden describir mejor usando una distribución normal con media $\mu_0 = \$30.00$ y desviación estándar $\sigma_0 = \$1.75$. Si una de estas máquinas despachadoras de jugo, ubicada en cierto edificio, muestra una utilidad promedio diaria de $\bar{x} = \$24.90$, durante los primeros 30 días con una desviación estándar de $s = \$2.10$, calcule

a) un estimado de Bayes de la utilidad promedio diaria verdadera para este edificio;

b) un intervalo bayesiano de 95% de μ para este edificio;

c) la probabilidad de que la utilidad promedio diaria de la máquina en este edificio sea de entre \$24.00 y \$26.00.

18.9 El departamento de matemáticas de una universidad grande diseña un examen de colocación para aplicarlo a los grupos de nuevo ingreso a primer año. Los miembros del departamento consideran que la calificación promedio para este examen variará de un grupo de primer año a otro. Esta variación de la calificación promedio del grupo se expresa de manera subjetiva mediante una distribución normal, con una media $\mu_0 = 72$ y una varianza $\sigma_0^2 = 5.76$.

a) ¿Qué probabilidad previa existe de que la calificación promedio real, que asigna el departamento para los alumnos de nuevo ingreso del siguiente año, caiga entre 71.8 y 73.4?

b) Construya un intervalo bayesiano de 95% para μ en el caso de que el examen se aplicara a una muestra aleatoria de 100 estudiantes de primer grado del siguiente grupo de nuevo ingreso y tuviera como resultado una calificación promedio de 70 con una varianza de 64.

c) ¿Qué probabilidad posterior debería asignar el departamento al evento del inciso *a*?

18.10 Suponga que en el ejemplo 18.7 la empresa de equipo eléctrico no tiene suficiente información previa respecto a la duración media poblacional que le permita suponer una distribución normal para μ. La empresa cree, sin embargo, que μ seguramente estará entre 770 y 830 horas, y considera que una aproximación bayesiana más realista sería suponer una distribución previa

$$\pi(\mu) = \frac{1}{60}, \quad 770 < \mu < 830.$$

Si una muestra aleatoria de 25 bombillas tiene una vida promedio de 780 horas, siga los pasos de la demostración del ejemplo 18.5 para encontrar la distribución posterior.

$$\pi(\mu \mid x_1, x_2, \ldots, x_{25}).$$

18.11 Suponga que el tiempo T antes de que falle cierta bisagra es una variable aleatoria exponencial con densidad de probabilidad

$$f(t) = \theta e^{-\theta t}, \quad t > 0.$$

Por experiencia, nos inclinamos a pensar que θ es un valor de una variable aleatoria exponencial con densidad de probabilidad

$$\pi(\theta) = 2e^{-2\theta}, \quad \theta > 0.$$

Si tenemos una muestra de n observaciones de T, demuestre que la distribución posterior de Θ es una distribución gamma

$$\alpha = n + 1 \quad \text{y} \quad \beta = \left(\sum_{i=1}^{n} t_i + 2 \right)^{-1}.$$

18.12 Suponga que una muestra consta de 5, 6, 6, 7, 5, 6, 4, 9 y 3, y 6 proviene de una población de Poisson con media λ. Suponga que el parámetro λ sigue una distribución gamma con parámetros (3, 2). Bajo la función de pérdida del cuadrado del error, calcule el estimado de Bayes de λ.

18.13 Una variable aleatoria X sigue una distribución binomial negativa con parámetros $k = 5$ y p, es decir, $b^*(x; 5, p)$. Además, se sabe que p sigue una distribución uniforme en el intervalo (0, 1). Calcule el estimado de Bayes de p bajo la función de pérdida del cuadrado del error.

18.14 Una variable aleatoria X sigue una distribución exponencial con media $1/\beta$. Suponga que la distribución previa de β es otra distribución exponencial con media 2.5. Determine el estimado de Bayes de β bajo la función de pérdida del error absoluto.

18.15 Una muestra aleatoria $X_1,..., X_n$ proviene de una población con distribución uniforme (véase la sección 6.1) con θ desconocida. Los datos se presentan a continuación:

0.13, 1.06, 1.65, 1.73, 0.95, 0.56, 2.14, 0.33, 1.22, 0.20, 1.55, 1.18, 0.71, 0.01, 0.42, 1.03, 0.43, 1.02, 0.83, 0.88

Suponga que la distribución previa de θ tiene la densidad

$$\pi(\theta) = \begin{cases} \frac{1}{\theta^2}, & \theta > 1, \\ 0, & \theta \leq 1. \end{cases}$$

Determine el estimador de Bayes bajo la función de pérdida del error absoluto.

Bibliografía

1 Bartlett, M. S. y Kendall, D. G. (1946). "The Statistical Analysis of Variance Heterogeneity and Logarithmic Transformation", *Journal of the Royal Statistical Society*, Ser. B. **8**, 128-138.

2 Bowker, A. H. y Lieberman, G. J. (1972). *Engineering Statistics*, 2.a, ed. Upper Saddle River, N.J.: Prentice Hall.

3 Box, G. E. P. (1988). "Signal to Noise Ratios, Performance Criteria and Transformations (with discussion)", *Technometrics*, **30**, 1-17.

4 Box, G. E. P. y Fung, C. A. (1986). "Studies in Quality Improvement: Minimizing Transmitted Variation by Parameter Design", Informe 8. University of Wisconsin-Madison, Center for Quality and Productivity Improvement.

5 Box, G. E. P., Hunter, W. G. y Hunter, J. S. (1978). *Statistics for Experimenters*. Nueva York: John Wiley & Sons.

6 Brownlee, K. A. (1984). *Statistical Theory and Methodology: In Science and Engineering*, 2.a, ed., Nueva York: John Wiley & Sons.

7 Carroll, R. J. y Ruppert, D. (1988). *Transformation and Weighting in Regression*. Nueva York: Chapman y Hall.

8 Chatterjee, S., Hadi, A. S. y Price, B. (1999). *Regression Analysis by Example*, 3.a, ed., Nueva York: John Wiley & Sons.

9 Cook, R. D. y Weisberg, S. (1982). *Residuals and Influence in Regression*. Nueva York: Chapman y Hall.

10 Daniel, C. y Wood, F. S. (1999). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2.a, ed., Nueva York: John Wiley & Sons.

11 Daniel, W. W. (1989). *Applied Nonparametric Statistics*, 2.a, ed. Belmont, Calif.: Wadsworth Publishing Company.

12 Devore, J. L. (2003). *Probability and Statistics for Engineering and the Sciences*, 6.a, ed., Belmont, Calif: Duxbury Press.

13 Dixon, W. J. (1983). *Introduction to Statistical Analysis*, 4.a, ed., Nueva York: McGraw-Hill.

14 Draper, N. R. y Smith, H. (1998). *Applied Regression Analysis*, 3.a, ed., Nueva York: John Wiley & Sons.

15 Duncan, A. (1986). *Quality Control and Industrial Statistics*, 5.a, ed., Homewood, Ill.: Irwin.

16 Dyer, D. D., y Keating, J. P. (1980). "On the Determination of Critical Values for Bartlett's Test", *Journal of the American Statistical Association*, **75**, 313-319.

17 Ewan, W. D. y Kemp, K. W. (1960). "Sampling Inspection of Continuous Processes with No Autocorrelation between Successive Results", *Biometrika*, **47**, 363-380.

18 Geary, R. C. (1947). "Testing for Normality", *Biometrika*, **34**, 209-242.

19 Gunst, R. F. y Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*. Nueva York: Marcel Dekker.

20 Guttman, I., Wilks, S. S. y Hunter, J. S. (1971). *Introductory Engineering Statistics*. Nueva York: John Wiley & Sons.

21 Harville, D. A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", *Journal of the American Statistical Association*, **72**, 320-338.

22 Hicks, C. R. y Turner, K. V. (1999). *Fundamental Concepts in the Design of Experiments*, 5.a, ed., Oxford: Oxford University Press.

23 Hoaglin, D. C., Mosteller, F. y Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance*. Nueva York: John Wiley & Sons.

24 Hocking, R. R. (1976). "The Analysis and Selection of Variables in Linear Regression", *Biometrics*, **32**, 1-49.

25 Hodges, J. L. y Lehmann, E. L. (2005). *Basic Concepts of Probability and Statistics*, 2.a, ed. Filadelfia: Society for Industrial and Applied Mathematics.

26 Hoerl, A. E. y Wennard, R. W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems", *Technometrics*, **12**, 55-67.

27 Hogg, R. V. y Ledolter, J. (1992). *Applied Statistics for Engineers and Physical Scientists*, 2.a, ed., Upper Saddle River, N.J.: Prentice Hall.

28 Hogg, R. V., McKean, J. W. y Craig, A. (2005). *Introduction to Mathematical Statistics*, 6.a, ed., Upper Saddle River, N.J.: Prentice Hall.

29 Hollander, M. y Wolfe, D. (1999). *Nonparametric Statistical Methods*. Nueva York: John Wiley & Sons.

30 Johnson, N. L. y Leone, F. C. (1977). *Statistics and Experimental Design: In Engineering and the Physical Sciences*, 2.a, ed., Vols. I y II, Nueva York: John Wiley & Sons.

31 Kackar, R. (1985). "Off-Line Quality Control, Parameter Design, and the Taguchi Methods", *Journal of Quality Technology*, **17**, 176-188.

32 Koopmans, L. H. (1987). *An Introduction to Contemporary Statistics*, 2.a, ed., Boston: Duxbury Press.

33 Kutner, M. H., Nachtsheim, C. J., Neter, J. y Li, W. (2004). *Applied Linear Regression Models*, 5.a, ed., Nueva York: McGraw-Hill/Irwin.

34 Larsen, R. J. y Morris, M. L. (2000). *An Introduction to Mathematical Statistics and Its Applications*, 3.a, ed., Upper Saddle River, N.J.: Prentice Hall.

35 Lehmann, E. L. y D'Abrera, H. J. M. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, ed. rev., Upper Saddle River, N.J.: Prentice Hall.

36 Lentner, M. y Bishop, T. (1986). *Design and Analysis of Experiments*, 2.a, ed., Blacksburg, Va.: Valley Book Co.

37 Mallows, C. L. (1973). "Some Comments on *Cp*", *Technometrics*, **15**, 661-675.

38 McClave, J. T., Dietrich, F. H. y Sincich, T. (1997). *Statistics*, 7.a, ed. Upper Saddle River, N.J.: Prentice Hall.

39 Montgomery, D. C. (2008a). *Design and Analysis of Experiments*, 7.a, ed., Nueva York: John Wiley & Sons.

40 Montgomery, D. C. (2008b). *Introduction to Statistical Quality Control*, 6.a, ed., Nueva York: John Wiley & Sons.

41 Mosteller, F. y Tukey, J. (1977). *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley Publishing Co.

42 Myers, R. H. (1990). *Classical and Modern Regression with Applications*, 2.a, ed., Boston: Duxbury Press.

43 Myers, R. H., Khuri, A. I. y Vining, G. G. (1992). "Response Surface Alternatives to the Taguchi Robust Parameter Design Approach", *The American Statistician*, **46**, 131-139.

44 Myers, R. H., Montgomery, D. C. y Anderson-Cook, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3.a, ed., Nueva York: John Wiley & Sons.

45 Myers, R. H., Montgomery, D. C., Vining, G. G. y Robinson, T. J. (2008). *Generalized Linear Models with Applications in Engineering and the Sciences*, 2.a, ed., Nueva York: John Wiley & Sons.

46 Noether, G. E. (1976). *Introduction to Statistics: A Nonparametric Approach*, 2.a, ed., Boston: Houghton Mifflin Company.

47 Olkin, I., Gleser, L. J. y Derman, C. (1994). *Probability Models and Applications*, 2.a, ed., Nueva York: Prentice Hall.

48 Ott, R. L. y Longnecker, M. T. (2000). *An Introduction to Statistical Methods and Data Analysis*, 5.a, ed., Boston: Duxbury Press.

49 Pacansky, J., England, C. D. y Wattman, R. (1986). "Infrared Spectroscopic Studies of Poly (perfluoropropyleneoxide) on Gold Substrate: A Classical Dispersion Analysis for the Refractive Index". *Applied Spectroscopy*, **40**, 8-16.

50 Plackett, R. L. y Burman, J. P. (1946). "The Design of Multifactor Experiments", *Biometrika*, **33**, 305-325.

51 Ross, S. M. (2002). *Introduction to Probability Models*, 9.a, ed., Nueva York: Academic Press, Inc.

52 Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components", *Biometrics*, **2**, 110-114.

53 Schilling, E. G. y Nelson, P. R. (1976). "The Effect of Nonnormality on the Control Limits of *X* Charts", *Journal of Quality Technology*, **8**, 347-373.

54 Schmidt, S. R. y Launsby, R. G. (1991). *Understanding Industrial Designed Experiments*. Colorado Springs, Col. Air Academy Press.

55 Shoemaker, A. C., Tsui, K.-L. y Wu, C. F. J. (1991). "Economical Experimentation Methods for Robust Parameter Design", *Technometrics*, **33**, 415-428.

56 Snedecor, G. W. y Cochran, W. G. (1989). *Statistical Methods*, 8a ed., Allies, Iowa: The Iowa State University Press.

57 Steel, R. G. D., Torrie, J. H. y Dickey, D. A. (1996). *Principles and Procedures of Statistics: A Biometrical Approach*, 3.a, ed., Nueva York: McGraw-Hill.

58 Taguchi, G. (1991). *Introduction to Quality Engineering*. White Plains, N.Y.: Unipub/Kraus International.

59 Taguchi, G. y Wu, Y. (1985). *Introduction to Off-Line Quality Control*. Nagoya, Japan: Central Japan Quality Control Association.

60 Thompson, W. O. y Cady, F. B. (1973). *Proceedings of the University of Kentucky Conference on Regression with a Large Number of Predictor Variables*. Lexington, Ken.: University of Kentucky Press.

61 Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley Publishing Co.

62 Vining, G. G. y Myers, R. H. (1990). "Combining Taguchi and Response Surface Philosophies: A Dual Response Approach", *Journal of Quality Technology*, **22**, 38-45.

63 Welch, W. J., Yu, T. K., Kang, S. M. y Sacks, J. (1990). "Computer Experiments for Quality Control by Parameter Design", *Journal of Quality Technology*, **22**, 15-22.

64 Winer, B. J. (1991). *Statistical Principles in Experimental Design*, 3.a, ed., Nueva York: McGraw-Hill.

Apéndice A
Tablas y demostraciones estadísticas

Tabla A.1 Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
1	**0**	0.9000	0.8000	0.7500	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	**0**	0.8100	0.6400	0.5625	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100
	1	0.9900	0.9600	0.9375	0.9100	0.8400	0.7500	0.6400	0.5100	0.3600	0.1900
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	**0**	0.7290	0.5120	0.4219	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010
	1	0.9720	0.8960	0.8438	0.7840	0.6480	0.5000	0.3520	0.2160	0.1040	0.0280
	2	0.9990	0.9920	0.9844	0.9730	0.9360	0.8750	0.7840	0.6570	0.4880	0.2710
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	**0**	0.6561	0.4096	0.3164	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001
	1	0.9477	0.8192	0.7383	0.6517	0.4752	0.3125	0.1792	0.0837	0.0272	0.0037
	2	0.9963	0.9728	0.9492	0.9163	0.8208	0.6875	0.5248	0.3483	0.1808	0.0523
	3	0.9999	0.9984	0.9961	0.9919	0.9744	0.9375	0.8704	0.7599	0.5904	0.3439
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	**0**	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0003	0.0000
	1	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0067	0.0005
	2	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.0579	0.0086
	3	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.2627	0.0815
	4	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.6723	0.4095
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	**0**	0.5314	0.2621	0.1780	0.1176	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000
	1	0.8857	0.6554	0.5339	0.4202	0.2333	0.1094	0.0410	0.0109	0.0016	0.0001
	2	0.9842	0.9011	0.8306	0.7443	0.5443	0.3438	0.1792	0.0705	0.0170	0.0013
	3	0.9987	0.9830	0.9624	0.9295	0.8208	0.6563	0.4557	0.2557	0.0989	0.0159
	4	0.9999	0.9984	0.9954	0.9891	0.9590	0.8906	0.7667	0.5798	0.3446	0.1143
	5	1.0000	0.9999	0.9998	0.9993	0.9959	0.9844	0.9533	0.8824	0.7379	0.4686
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	**0**	0.4783	0.2097	0.1335	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	
	1	0.8503	0.5767	0.4449	0.3294	0.1586	0.0625	0.0188	0.0038	0.0004	0.0000
	2	0.9743	0.8520	0.7564	0.6471	0.4199	0.2266	0.0963	0.0288	0.0047	0.0002
	3	0.9973	0.9667	0.9294	0.8740	0.7102	0.5000	0.2898	0.1260	0.0333	0.0027
	4	0.9998	0.9953	0.9871	0.9712	0.9037	0.7734	0.5801	0.3529	0.1480	0.0257
	5	1.0000	0.9996	0.9987	0.9962	0.9812	0.9375	0.8414	0.6706	0.4233	0.1497
	6		1.0000	0.9999	0.9998	0.9984	0.9922	0.9720	0.9176	0.7903	0.5217
	7			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (*continuación*) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
8	**0**	0.4305	0.1678	0.1001	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	
	1	0.8131	0.5033	0.3671	0.2553	0.1064	0.0352	0.0085	0.0013	0.0001	
	2	0.9619	0.7969	0.6785	0.5518	0.3154	0.1445	0.0498	0.0113	0.0012	0.0000
	3	0.9950	0.9437	0.8862	0.8059	0.5941	0.3633	0.1737	0.0580	0.0104	0.0004
	4	0.9996	0.9896	0.9727	0.9420	0.8263	0.6367	0.4059	0.1941	0.0563	0.0050
	5	1.0000	0.9988	0.9958	0.9887	0.9502	0.8555	0.6846	0.4482	0.2031	0.0381
	6		0.9999	0.9996	0.9987	0.9915	0.9648	0.8936	0.7447	0.4967	0.1869
	7		1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8322	0.5695
	8				1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	**0**	0.3874	0.1342	0.0751	0.0404	0.0101	0.0020	0.0003	0.0000		
	1	0.7748	0.4362	0.3003	0.1960	0.0705	0.0195	0.0038	0.0004	0.0000	
	2	0.9470	0.7382	0.6007	0.4628	0.2318	0.0898	0.0250	0.0043	0.0003	0.0000
	3	0.9917	0.9144	0.8343	0.7297	0.4826	0.2539	0.0994	0.0253	0.0031	0.0001
	4	0.9991	0.9804	0.9511	0.9012	0.7334	0.5000	0.2666	0.0988	0.0196	0.0009
	5	0.9999	0.9969	0.9900	0.9747	0.9006	0.7461	0.5174	0.2703	0.0856	0.0083
	6	1.0000	0.9997	0.9987	0.9957	0.9750	0.9102	0.7682	0.5372	0.2618	0.0530
	7		1.0000	0.9999	0.9996	0.9962	0.9805	0.9295	0.8040	0.5638	0.2252
	8			1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.8658	0.6126
	9					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	**0**	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000		
	1	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	
	2	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0001	
	3	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0009	0.0000
	4	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0064	0.0001
	5	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0328	0.0016
	6	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.1209	0.0128
	7		0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.3222	0.0702
	8		1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.6242	0.2639
	9				1.0000	0.9999	0.9990	0.9940	0.9718	0.8926	0.6513
	10					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	**0**	0.3138	0.0859	0.0422	0.0198	0.0036	0.0005	0.0000			
	1	0.6974	0.3221	0.1971	0.1130	0.0302	0.0059	0.0007	0.0000		
	2	0.9104	0.6174	0.4552	0.3127	0.1189	0.0327	0.0059	0.0006	0.0000	
	3	0.9815	0.8389	0.7133	0.5696	0.2963	0.1133	0.0293	0.0043	0.0002	
	4	0.9972	0.9496	0.8854	0.7897	0.5328	0.2744	0.0994	0.0216	0.0020	0.0000
	5	0.9997	0.9883	0.9657	0.9218	0.7535	0.5000	0.2465	0.0782	0.0117	0.0003
	6	1.0000	0.9980	0.9924	0.9784	0.9006	0.7256	0.4672	0.2103	0.0504	0.0028
	7		0.9998	0.9988	0.9957	0.9707	0.8867	0.7037	0.4304	0.1611	0.0185
	8		1.0000	0.9999	0.9994	0.9941	0.9673	0.8811	0.6873	0.3826	0.0896
	9			1.0000	1.0000	0.9993	0.9941	0.9698	0.8870	0.6779	0.3026
	10					1.0000	0.9995	0.9964	0.9802	0.9141	0.6862
	11						1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (*continuación*) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
12	**0**	0.2824	0.0687	0.0317	0.0138	0.0022	0.0002	0.0000			
	1	0.6590	0.2749	0.1584	0.0850	0.0196	0.0032	0.0003	0.0000		
	2	0.8891	0.5583	0.3907	0.2528	0.0834	0.0193	0.0028	0.0002	0.0000	
	3	0.9744	0.7946	0.6488	0.4925	0.2253	0.0730	0.0153	0.0017	0.0001	
	4	0.9957	0.9274	0.8424	0.7237	0.4382	0.1938	0.0573	0.0095	0.0006	0.0000
	5	0.9995	0.9806	0.9456	0.8822	0.6652	0.3872	0.1582	0.0386	0.0039	0.0001
	6	0.9999	0.9961	0.9857	0.9614	0.8418	0.6128	0.3348	0.1178	0.0194	0.0005
	7	1.0000	0.9994	0.9972	0.9905	0.9427	0.8062	0.5618	0.2763	0.0726	0.0043
	8		0.9999	0.9996	0.9983	0.9847	0.9270	0.7747	0.5075	0.2054	0.0256
	9		1.0000	1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.4417	0.1109
	10				1.0000	0.9997	0.9968	0.9804	0.9150	0.7251	0.3410
	11					1.0000	0.9998	0.9978	0.9862	0.9313	0.7176
	12						1.0000	1.0000	1.0000	1.0000	1.0000
13	**0**	0.2542	0.0550	0.0238	0.0097	0.0013	0.0001	0.0000			
	1	0.6213	0.2336	0.1267	0.0637	0.0126	0.0017	0.0001	0.0000		
	2	0.8661	0.5017	0.3326	0.2025	0.0579	0.0112	0.0013	0.0001		
	3	0.9658	0.7473	0.5843	0.4206	0.1686	0.0461	0.0078	0.0007	0.0000	
	4	0.9935	0.9009	0.7940	0.6543	0.3530	0.1334	0.0321	0.0040	0.0002	
	5	0.9991	0.9700	0.9198	0.8346	0.5744	0.2905	0.0977	0.0182	0.0012	0.0000
	6	0.9999	0.9930	0.9757	0.9376	0.7712	0.5000	0.2288	0.0624	0.0070	0.0001
	7	1.0000	0.9988	0.9944	0.9818	0.9023	0.7095	0.4256	0.1654	0.0300	0.0009
	8		0.9998	0.9990	0.9960	0.9679	0.8666	0.6470	0.3457	0.0991	0.0065
	9		1.0000	0.9999	0.9993	0.9922	0.9539	0.8314	0.5794	0.2527	0.0342
	10			1.0000	0.9999	0.9987	0.9888	0.9421	0.7975	0.4983	0.1339
	11				1.0000	0.9999	0.9983	0.9874	0.9363	0.7664	0.3787
	12					1.0000	0.9999	0.9987	0.9903	0.9450	0.7458
	13						1.0000	1.0000	1.0000	1.0000	1.0000
14	**0**	0.2288	0.0440	0.0178	0.0068	0.0008	0.0001	0.0000			
	1	0.5846	0.1979	0.1010	0.0475	0.0081	0.0009	0.0001			
	2	0.8416	0.4481	0.2811	0.1608	0.0398	0.0065	0.0006	0.0000		
	3	0.9559	0.6982	0.5213	0.3552	0.1243	0.0287	0.0039	0.0002		
	4	0.9908	0.8702	0.7415	0.5842	0.2793	0.0898	0.0175	0.0017	0.0000	
	5	0.9985	0.9561	0.8883	0.7805	0.4859	0.2120	0.0583	0.0083	0.0004	
	6	0.9998	0.9884	0.9617	0.9067	0.6925	0.3953	0.1501	0.0315	0.0024	0.0000
	7	1.0000	0.9976	0.9897	0.9685	0.8499	0.6047	0.3075	0.0933	0.0116	0.0002
	8		0.9996	0.9978	0.9917	0.9417	0.7880	0.5141	0.2195	0.0439	0.0015
	9		1.0000	0.9997	0.9983	0.9825	0.9102	0.7207	0.4158	0.1298	0.0092
	10			1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.3018	0.0441
	11				1.0000	0.9994	0.9935	0.9602	0.8392	0.5519	0.1584
	12					0.9999	0.9991	0.9919	0.9525	0.8021	0.4154
	13					1.0000	0.9999	0.9992	0.9932	0.9560	0.7712
	14						1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (*continuación*) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
15	**0**	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000				
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000			
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000		
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001		
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0000	
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0001	
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0008	
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0042	0.0000
	8		0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0181	0.0003
	9		0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.0611	0.0022
	10		1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.1642	0.0127
	11			1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.3518	0.0556
	12				1.0000	0.9997	0.9963	0.9729	0.8732	0.6020	0.1841
	13					1.0000	0.9995	0.9948	0.9647	0.8329	0.4510
	14						1.0000	0.9995	0.9953	0.9648	0.7941
	15							1.0000	1.0000	1.0000	1.0000
16	**0**	0.1853	0.0281	0.0100	0.0033	0.0003	0.0000				
	1	0.5147	0.1407	0.0635	0.0261	0.0033	0.0003	0.0000			
	2	0.7892	0.3518	0.1971	0.0994	0.0183	0.0021	0.0001			
	3	0.9316	0.5981	0.4050	0.2459	0.0651	0.0106	0.0009	0.0000		
	4	0.9830	0.7982	0.6302	0.4499	0.1666	0.0384	0.0049	0.0003		
	5	0.9967	0.9183	0.8103	0.6598	0.3288	0.1051	0.0191	0.0016	0.0000	
	6	0.9995	0.9733	0.9204	0.8247	0.5272	0.2272	0.0583	0.0071	0.0002	
	7	0.9999	0.9930	0.9729	0.9256	0.7161	0.4018	0.1423	0.0257	0.0015	0.0000
	8	1.0000	0.9985	0.9925	0.9743	0.8577	0.5982	0.2839	0.0744	0.0070	0.0001
	9		0.9998	0.9984	0.9929	0.9417	0.7728	0.4728	0.1753	0.0267	0.0005
	10		1.0000	0.9997	0.9984	0.9809	0.8949	0.6712	0.3402	0.0817	0.0033
	11			1.0000	0.9997	0.9951	0.9616	0.8334	0.5501	0.2018	0.0170
	12				1.0000	0.9991	0.9894	0.9349	0.7541	0.4019	0.0684
	13					0.9999	0.9979	0.9817	0.9006	0.6482	0.2108
	14					1.0000	0.9997	0.9967	0.9739	0.8593	0.4853
	15						1.0000	0.9997	0.9967	0.9719	0.8147
	16							1.0000	1.0000	1.0000	1.0000

Tabla A.1 (*continuación*) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
17	**0**	0.1668	0.0225	0.0075	0.0023	0.0002	0.0000				
	1	0.4818	0.1182	0.0501	0.0193	0.0021	0.0001	0.0000			
	2	0.7618	0.3096	0.1637	0.0774	0.0123	0.0012	0.0001			
	3	0.9174	0.5489	0.3530	0.2019	0.0464	0.0064	0.0005	0.0000		
	4	0.9779	0.7582	0.5739	0.3887	0.1260	0.0245	0.0025	0.0001		
	5	0.9953	0.8943	0.7653	0.5968	0.2639	0.0717	0.0106	0.0007	0.0000	
	6	0.9992	0.9623	0.8929	0.7752	0.4478	0.1662	0.0348	0.0032	0.0001	
	7	0.9999	0.9891	0.9598	0.8954	0.6405	0.3145	0.0919	0.0127	0.0005	
	8	1.0000	0.9974	0.9876	0.9597	0.8011	0.5000	0.1989	0.0403	0.0026	0.0000
	9		0.9995	0.9969	0.9873	0.9081	0.6855	0.3595	0.1046	0.0109	0.0001
	10		0.9999	0.9994	0.9968	0.9652	0.8338	0.5522	0.2248	0.0377	0.0008
	11		1.0000	0.9999	0.9993	0.9894	0.9283	0.7361	0.4032	0.1057	0.0047
	12			1.0000	0.9999	0.9975	0.9755	0.8740	0.6113	0.2418	0.0221
	13				1.0000	0.9995	0.9936	0.9536	0.7981	0.4511	0.0826
	14					0.9999	0.9988	0.9877	0.9226	0.6904	0.2382
	15					1.0000	0.9999	0.9979	0.9807	0.8818	0.5182
	16						1.0000	0.9998	0.9977	0.9775	0.8332
	17							1.0000	1.0000	1.0000	1.0000
18	**0**	0.1501	0.0180	0.0056	0.0016	0.0001	0.0000				
	1	0.4503	0.0991	0.0395	0.0142	0.0013	0.0001				
	2	0.7338	0.2713	0.1353	0.0600	0.0082	0.0007	0.0000			
	3	0.9018	0.5010	0.3057	0.1646	0.0328	0.0038	0.0002			
	4	0.9718	0.7164	0.5187	0.3327	0.0942	0.0154	0.0013	0.0000		
	5	0.9936	0.8671	0.7175	0.5344	0.2088	0.0481	0.0058	0.0003		
	6	0.9988	0.9487	0.8610	0.7217	0.3743	0.1189	0.0203	0.0014	0.0000	
	7	0.9998	0.9837	0.9431	0.8593	0.5634	0.2403	0.0576	0.0061	0.0002	
	8	1.0000	0.9957	0.9807	0.9404	0.7368	0.4073	0.1347	0.0210	0.0009	
	9		0.9991	0.9946	0.9790	0.8653	0.5927	0.2632	0.0596	0.0043	0.0000
	10		0.9998	0.9988	0.9939	0.9424	0.7597	0.4366	0.1407	0.0163	0.0002
	11		1.0000	0.9998	0.9986	0.9797	0.8811	0.6257	0.2783	0.0513	0.0012
	12			1.0000	0.9997	0.9942	0.9519	0.7912	0.4656	0.1329	0.0064
	13				1.0000	0.9987	0.9846	0.9058	0.6673	0.2836	0.0282
	14					0.9998	0.9962	0.9672	0.8354	0.4990	0.0982
	15					1.0000	0.9993	0.9918	0.9400	0.7287	0.2662
	16						0.9999	0.9987	0.9858	0.9009	0.5497
	17						1.0000	0.9999	0.9984	0.9820	0.8499
	18							1.0000	1.0000	1.0000	1.0000

Tabla A.1 (*continuación*) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
19	**0**	0.1351	0.0144	0.0042	0.0011	0.0001					
	1	0.4203	0.0829	0.0310	0.0104	0.0008	0.0000				
	2	0.7054	0.2369	0.1113	0.0462	0.0055	0.0004	0.0000			
	3	0.8850	0.4551	0.2631	0.1332	0.0230	0.0022	0.0001			
	4	0.9648	0.6733	0.4654	0.2822	0.0696	0.0096	0.0006	0.0000		
	5	0.9914	0.8369	0.6678	0.4739	0.1629	0.0318	0.0031	0.0001		
	6	0.9983	0.9324	0.8251	0.6655	0.3081	0.0835	0.0116	0.0006		
	7	0.9997	0.9767	0.9225	0.8180	0.4878	0.1796	0.0352	0.0028	0.0000	
	8	1.0000	0.9933	0.9713	0.9161	0.6675	0.3238	0.0885	0.0105	0.0003	
	9		0.9984	0.9911	0.9674	0.8139	0.5000	0.1861	0.0326	0.0016	
	10		0.9997	0.9977	0.9895	0.9115	0.6762	0.3325	0.0839	0.0067	0.0000
	11		1.0000	0.9995	0.9972	0.9648	0.8204	0.5122	0.1820	0.0233	0.0003
	12			0.9999	0.9994	0.9884	0.9165	0.6919	0.3345	0.0676	0.0017
	13			1.0000	0.9999	0.9969	0.9682	0.8371	0.5261	0.1631	0.0086
	14				1.0000	0.9994	0.9904	0.9304	0.7178	0.3267	0.0352
	15					0.9999	0.9978	0.9770	0.8668	0.5449	0.1150
	16					1.0000	0.9996	0.9945	0.9538	0.7631	0.2946
	17						1.0000	0.9992	0.9896	0.9171	0.5797
	18							0.9999	0.9989	0.9856	0.8649
	19							1.0000	1.0000	1.0000	1.0000
20	**0**	0.1216	0.0115	0.0032	0.0008	0.0000					
	1	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000				
	2	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002				
	3	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000			
	4	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003			
	5	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000		
	6	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003		
	7	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0000	
	8	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0001	
	9	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0006	
	10		0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0026	0.0000
	11		0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0100	0.0001
	12		1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.0321	0.0004
	13			1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.0867	0.0024
	14				1.0000	0.9984	0.9793	0.8744	0.5836	0.1958	0.0113
	15					0.9997	0.9941	0.9490	0.7625	0.3704	0.0432
	16					1.0000	0.9987	0.9840	0.8929	0.5886	0.1330
	17						0.9998	0.9964	0.9645	0.7939	0.3231
	18						1.0000	0.9995	0.9924	0.9308	0.6083
	19							1.0000	0.9992	0.9885	0.8784
	20								1.0000	1.0000	1.0000

Tabla A.2 Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x; \mu)$

	μ								
r	**0.1**	**0.2**	**0.30**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865
4		1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977
5				1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
6							1.0000	1.0000	1.0000

	μ								
r	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**
0	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160
6	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8		1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9			1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10				0.9999	0.9997	0.9990	0.9972	0.9933	0.9863
11				1.0000	0.9999	0.9997	0.9991	0.9976	0.9945
12					1.0000	0.9999	0.9997	0.9992	0.9980
13						1.0000	0.9999	0.9997	0.9993
14							1.0000	0.9999	0.9998
15								1.0000	0.9999
16									1.0000

Tabla A.2 (*continuación*) Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x; \mu)$

					μ				
r	**5.5**	**6.0**	**6.5**	**7.0**	**7.5**	**8.0**	**8.5**	**9.0**	**9.5**
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008
2	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149
4	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403
5	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885
6	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649
7	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687
8	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453
11	0.9890	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520
12	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981
14	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665
16	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823
17	1.0000	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911
18		1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957
19			1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980
20						0.9999	0.9998	0.9996	0.9991
21						1.0000	0.9999	0.9998	0.9996
22							1.0000	0.9999	0.9999
23								1.0000	0.9999
24									1.0000

Tabla A.2 (*continuación*) Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x; \mu)$

	μ								
r	**10.0**	**11.0**	**12.0**	**13.0**	**14.0**	**15.0**	**16.0**	**17.0**	**18.0**
0	0.0000	0.0000	0.0000						
1	0.0005	0.0002	0.0001	0.0000	0.0000				
2	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000		
3	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000
4	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001
5	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003
6	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0010
7	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0029
8	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0071
9	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154
10	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304
11	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0549
12	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0917
13	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1426
14	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081
15	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867
16	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3751
17	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.4686
18	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.5622
19	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509
20	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.7307
21	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991
22	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551
23	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8989
24	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9317
25		0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9554
26		1.0000	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718
27			0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827
28			1.0000	0.9999	0.9997	0.9991	0.9978	0.9950	0.9897
29				1.0000	0.9999	0.9996	0.9989	0.9973	0.9941
30					0.9999	0.9998	0.9994	0.9986	0.9967
31					1.0000	0.9999	0.9997	0.9993	0.9982
32						1.0000	0.9999	0.9996	0.9990
33							0.9999	0.9998	0.9995
34							1.0000	0.9999	0.9998
35								1.0000	0.9999
36									0.9999
37									1.0000

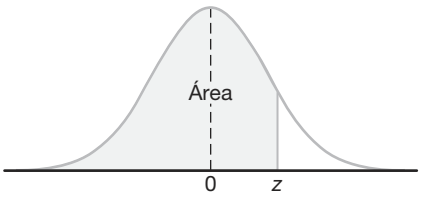


Tabla A.3 Áreas bajo la curva normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Tabla A.3 (*continuación*) Áreas bajo la curva normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

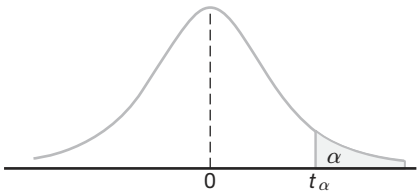


Tabla A.4 Valores críticos de la distribución *t*

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Tabla A.4 (*continuación*) Valores críticos de la distribución *t*

	α						
ν	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

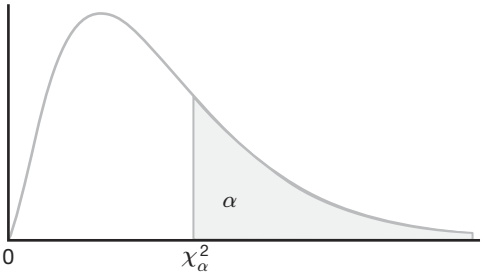


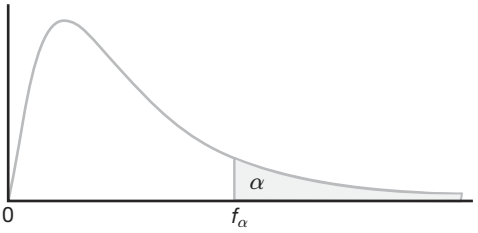
Tabla A.5 Valores críticos de la distribución chi cuadrada

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.041	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.878	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.66	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Tabla A.5 (*continuación*) Valores críticos de la distribución chi cuadrada

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.819	34.527
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.698
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.791
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.819
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.796
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.619
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.051
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.475
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.994	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.335	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.702
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.403
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.660
60	65.226	66.981	68.972	74.397	79.082	83.298	84.58	88.379	91.952	99.608

Tabla A.6 Valores críticos de la distribución F



	$f_{0.05}(v_1, v_2)$								
	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Reproducida de la tabla 18 de *Biometrika Tables for Statisticians*, Vol. I, con autorización de E.S. Pearson y Biometrika Trustees.

Tabla A.6 (*continuación*) Valores críticos de la distribución *F*

	$f_{0.05}(v_1, v_2)$									
	v_1									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Tabla A.6 (*continuación*) Valores críticos de la distribución *F*

	$f_{0.01}(v_1, v_2)$								
	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Tabla A.6 (*continuación*) Valores críticos de la distribución *F*

	$f_{0.01}(v_1, v_2)$									
	v_1									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabla A.7 Factores de tolerancia para distribuciones normales

	Intervalos bilaterales						Intervalos unilaterales					
	$\gamma = 0.05$			$\gamma = 0.01$			$\gamma = 0.05$			$\gamma = 0.01$		
	$1 - \alpha$			$1 - \alpha$			$1 - \alpha$			$1 - \alpha$		
n	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**
2	32.019	37.674	48.430	160.193	188.491	242.300	20.581	26.260	37.094	103.029	131.426	185.617
3	8.380	9.916	12.861	18.930	22.401	29.055	6.156	7.656	10.553	13.995	17.170	23.896
4	5.369	6.370	8.299	9.398	11.150	14.527	4.162	5.144	7.042	7.380	9.083	12.387
5	4.275	5.079	6.634	6.612	7.855	10.260	3.407	4.203	5.741	5.362	6.578	8.939
6	3.712	4.414	5.775	5.337	6.345	8.301	3.006	3.708	5.062	4.411	5.406	7.335
7	3.369	4.007	5.248	4.613	5.488	7.187	2.756	3.400	4.642	3.859	4.728	6.412
8	3.136	3.732	4.891	4.147	4.936	6.468	2.582	3.187	4.354	3.497	4.285	5.812
9	2.967	3.532	4.631	3.822	4.550	5.966	2.454	3.031	4.143	3.241	3.972	5.389
10	2.839	3.379	4.433	3.582	4.265	5.594	2.355	2.911	3.981	3.048	3.738	5.074
11	2.737	3.259	4.277	3.397	4.045	5.308	2.275	2.815	3.852	2.898	3.556	4.829
12	2.655	3.162	4.150	3.250	3.870	5.079	2.210	2.736	3.747	2.777	3.410	4.633
13	2.587	3.081	4.044	3.130	3.727	4.893	2.155	2.671	3.659	2.677	3.290	4.472
14	2.529	3.012	3.955	3.029	3.608	4.737	2.109	2.615	3.585	2.593	1.189	4.337
15	2.480	2.954	3.878	2.945	3.507	4.605	2.068	2.566	3.520	2.522	3.102	4.222
16	2.437	2.903	3.812	2.872	3.421	4.492	2.033	2.524	3.464	2.460	3.028	4.123
17	2.400	2.858	3.754	2.808	3.345	4.393	2.002	2.486	3.414	2.405	2.963	4.037
18	2.366	2.819	3.702	2.753	3.279	4.307	1.974	2.453	3.370	2.357	2.905	3.960
19	2.337	2.784	3.656	2.703	3.221	4.230	1.949	2.423	3.331	2.314	2.854	3.892
20	2.310	2.752	3.615	2.659	3.168	4.161	1.926	2.396	3.295	2.276	2.808	1.832
25	2.208	2.631	3.457	2.494	2.972	3.904	1.838	2.292	3.158	2.129	2.633	3.001
30	2.140	2.549	3.350	2.385	2.841	3.733	1.777	2.220	3.064	2.030	2.516	3.447
35	2.090	2.490	3.272	2.306	2.748	3.611	1.732	2.167	2.995	1.957	2.430	3.334
40	2.052	2.445	3.213	2.247	2.677	3.518	1.697	2.126	2.941	1.902	2.364	3.249
45	2.021	2.408	3.165	2.200	2.621	3.444	1.669	2.092	2.898	1.857	2.312	3.180
50	1.996	2.379	3.126	2.162	2.576	3.385	1.646	2.065	2.863	1.821	2.269	3.125
60	1.958	2.333	3.066	2.103	2.506	3.293	1.609	2.022	2.807	1.764	2.202	3.038
70	1.929	2.299	3.021	2.060	2.454	3.225	1.581	1.990	2.765	1.722	2.153	2.974
80	1.907	2.272	2.986	2.026	2.414	3.173	1.559	1.965	2.733	1.688	2.114	2.924
90	1.889	2.251	2.958	1.999	2.382	3.130	1.542	1.944	2.706	1.661	2.082	2.883
100	1.874	2.233	2.934	1.977	2.355	3.096	1.527	1.927	2.684	1.639	2.056	2.850
150	1.825	2.175	2.859	1.905	2.270	2.983	1.478	1.870	2.611	1.566	1.971	2.741
200	1.798	2.143	2.816	1.865	2.222	2.921	1.450	1.837	2.570	1.524	1.923	2.679
250	1.780	2.121	2.788	1.839	2.191	2.880	1.431	1.815	2.542	1.496	1.891	2.638
300	1.767	2.106	2.767	1.820	2.169	2.850	1.417	1.800	2.522	1.476	1.868	2.608
∞	1.645	1.960	2.576	1.645	1.960	2.576	1.282	1.645	2.326	1.282	1.645	2.326

Adaptada de C. Eisenhart, M. W. Hastay y W. A. Wallis, *Techniques of Statistical Analysis*, capítulo 2, McGraw-Hill Book Company, Nueva York, 1947. Se utiliza con autorización de McGraw-Hill Book Company.

Tabla A.8 Tamaño muestral para la prueba *t* de la media

		Nivel de la prueba *t*																			
Prueba unilateral		$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$				
Prueba bilateral		$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$				
$\beta = 0.1$		**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	***.5***
	0.05																				
	0.10																				
	0.15																				122
	0.20										139					99					70
	0.25					110					90				128	64			139	101	45
	0.30				134	78				115	63			119	90	45		122	97	71	32
	0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24
	0.40		115	97	77	45		101	85	66	37	117	84	68	51	26	101	70	55	40	19
	0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15
	0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13
	0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11
	0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9
	0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8
	0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8
	0.75	47	36	30	25	16	42	31	27	21	1	35	26	21	16	9	30	21	17	13	7
	0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6
	0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6
Valor de	**0.90**	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5
$\Delta = \lvert\delta\rvert/\sigma$	**0.95**	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5
	1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5
	1.1	24	19	16	14	9	21	16	14	12	8	18	13	11	9	6	15	11	9	7	
	1.2	21	16	14	12	8	18	14	12	10	7	15	12	10	8	5	13	10	8	6	
	1.3	18	15	13	11	8	16	13	11	9	6	13	10	9	7		11	8	7	6	
	1.4	16	13	12	10	7	14	11	10	9	6	12	9	8	7		10	8	7	5	
	1.5	15	12	11	9	7	13	10	9	8	6	11	8	7	6		9	7	6		
	1.6	13	11	10	8	6	12	10	9	7	5	10	8	7	6		8	6	6		
	1.7	12	10	9	8	6	11	9	8	7		9	7	6	5		8	6	5		
	1.8	12	10	9	8	6	10	8	7	7		8	7	6			7	6			
	1.9	11	9	8	7	6	10	8	7	6		8	6	6			7	5			
	2.0	10	8	8	7	5	9	7	7	6		7	6	5			6				
	2.1	10	8	7	7		8	7	6	6		7	6				6				
	2.2	9	8	7	6		8	7	6	5		7	6				6				
	2.3	9	7	7	6		8	6	6			6	5				5				
	2.4	8	7	7	6		7	6	6			6									
	2.5	8	7	6	6		7	6	6			6									
	3.0	7	6	6	5		6	5	5			5									
	3.5	6	5	5			5														
	4.0	6																			

Reproducida con autorización de O. L. Davies, ed., *Design and Analysis of Industrial Experiments*, Oliver & Boyd, Edimburgo, 1956.

Tabla A.9 Tamaño muestral para la prueba *t* de la diferencia entre dos medias

		Nivel de la prueba *t*																			
Prueba unilateral		$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$				
Prueba bilateral		$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$				
$\beta = 0.1$		**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**
	0.05																				137
	0.10															124					88
	0.15										123					87					61
	0.20					110					90					64				102	45
	0.25					85					70				100	50			108	78	35
	0.30				118	68				101	55			105	79	39		108	86	62	28
	0.35				96	55			106	82	45		106	86	64	32		88	70	51	23
	0.40			101	79	46		106	88	68	38		87	71	53	27	112	73	58	42	19
	0.45		101	85	67	39		90	74	58	32	104	74	60	45	23	89	61	49	36	16
	0.50		87	73	57	34	104	77	64	49	27	88	63	51	39	20	76	52	42	30	14
	0.55	100	75	63	50	29	90	66	55	43	24	76	55	44	34	17	66	45	36	26	12
	0.60	88	66	55	44	26	79	58	48	38	21	67	48	39	29	15	57	40	32	23	11
	0.65	77	58	49	39	23	70	51	43	33	19	59	42	34	26	14	50	35	28	21	10
	0.70	69	51	43	35	21	62	46	38	30	17	52	37	31	23	12	45	31	25	18	9
Valor de	**0.75**	62	46	39	31	19	55	41	34	27	15	47	34	27	21	11	40	28	22	16	8
$\Delta = \lvert\delta\rvert/\sigma$	**0.80**	55	42	35	28	17	50	37	31	24	14	42	30	25	19	10	36	25	20	15	7
	0.85	50	38	32	26	15	45	33	28	22	13	38	27	23	17	9	33	23	18	14	7
	0.90	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	12	6
	0.95	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5
	1.00	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5
	1.1	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4
	1.2	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4
	1.3	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4
	1.4	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3
	1.5	17	13	71	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5	
	1.6	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5	
	1.7	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4	
	1.8	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4	
	1.9	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4	
	2.0	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4	
	2.1	11	9	8	6	5	10	8	7	6	4	8	6	5	4		7	5	4	4	
	2.2	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3	
	2.3	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3		
	2.4	6	5	5	4	3	6	5	4	4	5	4	4	3	3		4	3			
	2.5	6	5	4	4		5	4	4	3	4	4	3				4				
	3.0																				
	3.5																				
	4.0																				

Reproducida con autorización de O. L. Davies, ed., *Design and Analysis of Industrial Experiments*, Oliver & Boyd, Edimburgo, 1956.

Tabla A.10 Valores críticos para la prueba de Bartlett

	$b_k(0.01; n)$								
	Número de poblaciones, k								
n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0.1411	0.1672							
4	0.2843	0.3165	0.3475	0.3729	0.3937	0.4110			
5	0.3984	0.4304	0.4607	0.4850	0.5046	0.5207	0.5343	0.5458	0.5558
6	0.4850	0.5149	0.5430	0.5653	0.5832	O.5978	0.6100	0.6204	0.6293
7	0.5512	0.5787	0.6045	0.6248	0.6410	0.6542	0.6652	0.6744	0.6824
8	0.6031	0.6282	0.6518	0.6704	0.6851	0.6970	0.7069	0.7153	0.7225
9	0.6445	0.6676	0.6892	0.7062	0.7197	0.7305	0.7395	0.7471	0.7536
10	0.6783	0.6996	0.7195	0.7352	0.7475	0.7575	0.7657	0.7726	0.7786
11	0.7063	0.7260	0.7445	0.7590	0.7703	0.7795	0.7871	0.7935	0.7990
12	0.7299	0.7483	0.7654	0.7789	0.7894	0.7980	0.8050	0.8109	0.8160
13	0.7501	0.7672	0.7832	0.7958	0.8056	0.8135	0.8201	0.8256	0.8303
14	0.7674	0.7835	0.7985	0.8103	0.8195	0.8269	0.8330	0.8382	0.8426
15	0.7825	0.7977	0.8118	0.8229	0.8315	0.8385	0.8443	0.8491	0.8532
16	0.7958	0.8101	0.8235	0.8339	0.8421	0.8486	0.8541	0.8586	0.8625
17	0.8076	0.8211	0.8338	0.8436	0.8514	0.8576	0.8627	0.8670	0.8707
18	0.8181	0.8309	0.8429	0.8523	0.8596	0.8655	0.8704	0.8745	0.8780
19	0.8275	0.8397	0.8512	0.8601	0.8670	0.8727	0.8773	0.8811	0.8845
20	0.8360	0.8476	0.8586	0.8671	0.8737	0.8791	0.8835	0.8871	0.8903
21	0.8437	0.8548	0.8653	0.8734	0.8797	0.8848	0.8890	0.8926	0.8956
22	0.8507	0.8614	0.8714	0.8791	0.8852	0.8901	0.8941	0.8975	0.9004
23	0.8571	0.8673	0.8769	0.8844	0.8902	0.8949	0.8988	0.9020	0.9047
24	0.8630	0.8728	0.8820	0.8892	0.8948	0.8993	0.9030	0.9061	0.9087
25	0.8684	0.8779	0.8867	0.8936	0.8990	0.9034	0.9069	0.9099	0.9124
26	0.8734	0.8825	0.8911	0.8977	0.9029	0.9071	0.9105	0.9134	0.9158
27	0.8781	0.8869	0.8951	0.9015	0.9065	0.9105	0.9138	0.9166	0.9190
28	0.8824	0.8909	0.8988	0.9050	0.9099	0.9138	0.9169	0.9196	0.9219
29	0.8864	0.8946	0.9023	0.9083	0.9130	0.9167	0.9198	0.9224	0.9246
30	0.8902	0.8981	0.9056	0.9114	0.9159	0.9195	0.9225	0.9250	0.9271
40	0.9175	0.9235	0.9291	0.9335	0.9370	0.9397	0.9420	0.9439	0.9455
50	0.9339	0.9387	0.9433	0.9468	0.9496	0.9518	0.9536	0.9551	0.9564
60	0.9449	0.9489	0.9527	0.9557	0.9580	0.9599	0.9614	0.9626	0.9637
80	0.9586	0.9617	0.9646	0.9668	0.9685	0.9699	0.9711	0.9720	0.9728
100	0.9669	0.9693	0.9716	0.9734	0.9748	0.9759	0.9769	0.9776	0.9783

Reproducida de D. D. Dyer y J. P. Keating, "On the Determination of Critical Values for Bartlett's Test", *J. Am. Stat. Assoc.*, **75**, 1980, con autorización del consejo de directores.

Tabla A.10 (*continuación*) Valores críticos para la prueba de Bartlett

	$b_k(0.05; n)$								
	Número de poblaciones, *k*								
n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0.3123	0.3058	0.3173	0.3299					
4	0.4780	0.4699	0.4803	0.4921	0.5028	0.5122	0.5204	0.5277	0.5341
5	0.5845	0.5762	0.5850	0.5952	0.6045	0.6126	0.6197	0.6260	0.6315
6	0.6563	0.6483	0.6559	0.6646	0.6727	0.6798	0.6860	0.6914	0.6961
7	0.7075	0.7000	0.7065	0.7142	0.7213	0.7275	0.7329	0.7376	0.7418
8	0.7456	0.7387	0.7444	0.7512	0.7574	0.7629	0.7677	0.7719	0.7757
9	0.7751	0.7686	0.7737	0.7798	0.7854	0.7903	0.7946	0.7984	0.8017
10	0.7984	0.7924	0.7970	0.8025	0.8076	0.8121	0.8160	0.8194	0.8224
11	0.8175	0.8118	0.8160	0.8210	0.8257	0.8298	0.8333	0.8365	0.8392
12	0.8332	0.8280	0.8317	0.8364	0.8407	0.8444	0.8477	0.8506	0.8531
13	0.8465	0.8415	0.8450	0.8493	0.8533	0.8568	0.8598	0.8625	0.8648
14	0.8578	0.8532	0.8564	0.8604	0.8641	0.8673	0.8701	0.8726	0.8748
15	0.8676	0.8632	0.8662	0.8699	0.8734	0.8764	0.8790	0.8814	0.8834
16	0.8761	0.8719	0.8747	0.8782	0.8815	0.8843	0.8868	0.8890	0.8909
17	0.8836	0.8796	0.8823	0.8856	0.8886	0.8913	0.8936	0.8957	0.8975
18	0.8902	0.8865	0.8890	0.8921	0.8949	0.8975	0.8997	0.9016	0.9033
19	0.8961	0.8926	0.8949	0.8979	0.9006	0.9030	0.9051	0.9069	0.9086
20	0.9015	0.8980	0.9003	0.9031	0.9057	0.9080	0.9100	0.9117	0.9132
21	0.9063	0.9030	0.9051	0.9078	0.9103	0.9124	0.9143	0.9160	0.9175
22	0.9106	0.9075	0.9095	0.9120	0.9144	0.9165	0.9183	0.9199	0.9213
23	0.9146	0.9116	0.9135	0.9159	0.9182	0.9202	0.9219	0.9235	0.9248
24	0.9182	0.9153	0.9172	0.9195	0.9217	0.9236	0.9253	0.9267	0.9280
25	0.9216	0.9187	0.9205	0.9228	0.9249	0.9267	0.9283	0.9297	0.9309
26	0.9246	0.9219	0.9236	0.9258	0.9278	0.9296	0.9311	0.9325	0.9336
27	0.9275	0.9249	0.9265	0.9286	0.9305	0.9322	0.9337	0.9350	0.9361
28	0.9301	0.9276	0.9292	0.9312	0.9330	0.9347	0.9361	0.9374	0.9385
29	0.9326	0.9301	0.9316	0.9336	0.9354	0.9370	0.9383	0.9396	0.9406
30	0.9348	0.9325	0.9340	0.9358	0.9376	0.9391	0.9404	0.9416	0.9426
40	0.9513	0.9495	0.9506	0.9520	0.9533	0.9545	0.9555	0.9564	0.9572
50	0.9612	0.9597	0.9606	0.9617	0.9628	0.9637	0.9645	0.9652	0.9658
60	0.9677	0.9665	0.9672	0.9681	0.9690	0.9698	0.9705	0.9710	0.9716
80	0.9758	0.9749	0.9754	0.9761	0.9768	0.9774	0.9779	0.9783	0.9787
100	0.9807	0.9799	0.9804	0.9809	0.9815	0.9819	0.9823	0.9827	0.9830

Tabla A.11 Valores críticos para la prueba de Cochran

$\alpha = 0.01$														
	n													
k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**17**	**37**	**145**	∞
2	0.9999	0.9950	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000
3	0.9933	0.9423	0.8831	0.8335	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Reproducida de C. Eisenhart, M. W. Hastay y W. A. Wallis, *Techniques of Statistical Analysis*, capítulo 15, McGraw-Hill Book Company, Nueva York, 1947. Utilizada con autorización de McGraw-Hill Book Company.

Tabla A.11 (*continuación*) Valores críticos para la prueba de Cochran

$\alpha = 0.05$														
	n													
k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**17**	**37**	**145**	∞
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6333	0.6167	0.6025	0.5466	0.4748	0.4031	0.3333
4	0.9065	0.7679	0.6841	0.6287	0.5895	0.5598	0.5365	0.5175	0.5017	0.4884	0.4366	0.3720	0.3093	0.2500
5	0.8412	0.6838	0.5981	0.5441	0.5065	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2513	0.2000
6	0.7808	0.6161	0.5321	0.4803	0.4447	0.4184	0.3980	0.3817	0.3682	0.3568	0.3135	0.2612	0.2119	0.1667
7	0.7271	0.5612	0.4800	0.4307	0.3974	0.3726	0.3535	0.3384	0.3259	0.3154	0.2756	0.2278	0.1833	0.1429
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250
9	0.6385	0.4775	0.4027	0.3584	0.3286	0.3067	0.2901	0.2768	0.2659	0.2568	0.2226	0.1820	0.1446	0.1111
10	6.6020	0.4450	0.3733	0.3311	0.3029	0.2823	0.2666	0.2541	0.2439	0.2353	0.2032	0.1655	0.1308	0.1000
12	0.5410	0.3924	0.3264	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833
15	0.4709	0.3346	0.2758	0.2419	0.2195	0.2034	0.1911	0.1815	0.1736	0.1671	0.1429	0.1144	0.0889	0.0667
20	0.3894	0.2705	0.2205	0.1921	0.1735	0.1602	0.1501	0.1422	0.1357	0.1303	0.1108	0.0879	0.0675	0.0500
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417
30	0.2929	0.1980	0.1593	0.1377	0.1237	0.1137	0.1061	0.1002	0.0958	0.0921	0.0771	0.0604	0.0457	0.0333
40	0.2370	0.1576	0.1259	0.1082	0.0968	0.0887	0.0827	0.0780	0.0745	0.0713	0.0595	0.0462	0.0347	0.0250
60	0.1737	0.1131	0.0895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0.0167
120	0.0998	0.0632	0.0495	0.0419	0.0371	0.0337	0.0312	0.0292	0.0279	0.0266	0.0218	0.0165	0.0120	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla A.12 Puntos porcentuales superiores de la distribución de rangos estudentizados: valores de $q(0.05; k, v)$

Grados de libertad, v	Número de tratamientos, k								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	18.0	27.0	32.8	37.2	40.5	43.1	15.1	47.1	49.1
2	6.09	5.33	9.80	10.89	11.73	12.43	13.03	13.54	13.99
3	4.50	5.91	6.83	7.51	8.04	8.47	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.06	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.35	5.59	5.80	5.99	6.15
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.65	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.05
17	2.98	3.62	4.02	4.31	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.83	4.96	5.07
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.90	5.01
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
30	2.89	3.48	3.84	4.11	4.30	4.46	4.60	4.72	4.83
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47

Tabla A.13 Rangos estudentizados significativos mínimos $r_p(0.05; p, v)$

	$\alpha = 0.05$								
	p								
v	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.68	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.342	3.397	3.435	3.462	3.48	3.493	3.501
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457
15	3.014	3.160	3.25	3.312	3.356	3.389	3.413	3.432	3.446
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

Condensada de H. L. Harter, "Critical Values for Duncan's New Multiple Range Test", *Biometrics*, **16**, núm. 4, 1960, con autorización del autor y del editor.

Tabla A.13 (*continuación*) Rangos estudentizados significativos mínimos $r_p(0.01; p, v)$

$\alpha = 0.01$									
	p								
v	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.261	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.512	6.677	6.740	6.756	6.756	6.756	6.756	6.756	6.756
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703
7	4.949	5.145	5.260	5.334	5.383	5.416	5.439	5.454	5.464
8	4.746	4.939	5.057	5.135	5.189	5.227	5.256	5.276	5.291
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142	5.160
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205

Tabla A.14 Valores de $d_{\alpha/2}(k, v)$ para comparaciones bilaterales entre k tratamientos y un control

$\alpha = 0.05$									
	k = número de medias de tratamiento (no incluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69

Reproducida de Charles W. Dunnett, "New Tables for Multiple Comparison with a Control", *Biometrics*, **20**, núm. 3, 1964, con autorización del autor y del editor.

Tabla A.14 (*continuación*) Valores de $d_{\alpha/2}(k, v)$ para comparaciones bilaterales entre k tratamientos y un control

$\alpha = 0.01$									
	k = número de medias de tratamiento (no incluye el control)								
v	1	2	3	4	5	6	7	8	9
5	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89
6	3.71	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28
7	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89
8	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62
9	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43
10	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28
11	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16
12	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07
13	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99
14	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93
15	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88
16	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83
17	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79
18	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75
19	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72
20	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69
24	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61
30	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52
40	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44
60	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37
120	2.62	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29
∞	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22

Tabla A.15 Valores de $d_\alpha(k, v)$ para comparaciones unilaterales entre k tratamientos y un control

	$\alpha = 0.05$								
	k = número de medias de tratamiento (no incluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30
6	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12
7	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01
8	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92
9	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86
10	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81
11	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77
12	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74
13	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71
14	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69
15	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67
16	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65
17	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64
18	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62
19	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61
20	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60
24	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57
30	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54
40	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51
60	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48
120	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45
∞	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42

Reproducida de Charles W. Dunnett, "A Multiple Comparison Procedure for Comparing Several Treatments with a Control", *J. Am. Stat. Assoc*., **50**, 1955, 1096-1121, con autorización del autor y del editor.

Tabla A.15 (*continuación*) Valores de $d_\alpha(k, v)$ para comparaciones unilaterales entre k tratamientos y un control

$\alpha = 0.01$									
	k = número de medias de tratamiento (no incluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Tabla A.16 Valores críticos para la prueba de rangos con signo

n	Unilateral $\alpha = 0.01$ Bilateral $\alpha = 0.02$	Unilateral $\alpha = 0.025$ Bilateral $\alpha = 0.05$	Unilateral $\alpha = 0.05$ Bilateral $\alpha = 0.1$
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11
11	7	11	14
12	10	14	17
13	13	17	21
14	16	21	26
15	20	25	30
16	24	30	36
17	28	35	41
18	33	40	47
19	38	46	54
20	43	52	60
21	49	59	68
22	56	66	75
23	62	73	83
24	69	81	92
25	77	90	101
26	85	98	110
27	93	107	120
28	102	117	130
29	111	127	141
30	120	137	152

Reproducida de F. Wilcoxon y R. A. Wilcox, *Some Rapid Approximate Statistical Procedures, American Cyanamid Company*, Pearl River, N. Y., 1964, con autorización de la American Cyanamid Company.

Tabla A.17 Valores críticos para la prueba de suma de rangos de Wilcoxon

Prueba de una cola con $\alpha = 0.001$ o prueba de dos colas con $\alpha = 0.002$

	n_2														
n_1	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1															
2															
3												0	0	0	0
4					0	0	0	1	1	1	2	2	3	3	3
5		0	0	1	1	2	2	3	3	4	5	5	6	7	7
6	0	1	2	2	3	4	4	5	6	7	8	9	10	11	12
7		2	3	3	5	6	7	8	9	10	11	13	14	15	16
8			5	5	6	8	9	11	12	14	15	17	18	20	21
9				7	8	10	12	14	15	17	19	21	23	25	26
10					10	12	14	17	19	21	23	25	27	29	32
11						15	17	20	22	24	27	29	32	34	37
12							20	23	25	28	31	34	37	40	42
13								26	29	32	35	38	42	45	48
14									32	36	39	43	46	50	54
15										40	43	47	51	55	59
16											48	52	56	60	65
17												57	61	66	70
18													66	71	76
19														77	82
20															88

Prueba de una cola con $\alpha = 0.01$ o prueba de dos colas con $\alpha = 0.02$

	n_2														
n_1	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																
2									0	0	0	0	0	0	1	1
3			0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6		3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			6	8	9	11	12	14	16	17	19	21	23	24	26	28
8				10	11	13	15	17	20	22	24	26	28	30	32	34
9					14	16	18	21	23	26	28	31	33	36	38	40
10						19	22	24	27	30	33	36	38	41	44	47
11							25	28	31	34	37	41	44	47	50	53
12								31	35	38	42	46	49	53	56	60
13									39	43	47	51	55	59	63	67
14										47	51	56	60	65	69	73
15											56	61	66	70	75	80
16												66	71	76	82	87
17													77	82	88	93
18														88	94	100
19															101	107
20																114

Basada en parte en las tablas 1, 3, 5 y 7 de D. Auble, "Extended Tables for the Mann-Whitney Statistic", *Bulletin of the Institute of Educational Research at Indiana University*, **1**, núm. 2, 1953, con autorización del director.

Tabla A.17 (*continuación*) Valores críticos para la prueba de suma de rangos de Wilcoxon

Prueba de una cola con $\alpha = 0.025$ o prueba de dos colas con $\alpha = 0.05$

	n_2																
n_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																	
2					0	0	0	0	1	1	1	1	1	2	2	2	2
3		0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	67	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

Prueba de una cola con $\alpha = 0.05$ o prueba de dos colas con $\alpha = 0.1$

	n_2																
n_1	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																	0	0
2			0	0	0	1	1	1	1	2	2	3	3	3	3	4	4	4
3	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						15	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130
20																		138

Tabla A.18 $P(V \leq v^*$ cuando H_0 es verdadera) en la prueba de rachas

	v^*								
(n_1, n_2)	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
(2, 3)	0.200	0.500	0.900	1.000					
(2, 4)	0.133	0.400	0.800	1.000					
(2, 5)	0.095	0.333	0.714	1.000					
(2, 6)	0.071	0.286	0.643	1.000					
(2, 7)	0.056	0.250	0.583	1.000					
(2, 8)	0.044	0.222	0.533	1.000					
(2, 9)	0.036	0.200	0.491	1.000					
(2, 10)	0.030	0.182	0.455	1.000					
(3, 3)	0.100	0.300	0.700	0.900	1.000				
(3, 4)	0.057	0.200	0.543	0.800	0.971	1.000			
(3, 5)	0.036	0.143	0.429	0.714	0.929	1.000			
(3, 6)	0.024	0.107	0.345	0.643	0.881	1.000			
(3, 7)	0.017	0.083	0.283	0.583	0.833	1.000			
(3, 8)	0.012	0.067	0.236	0.533	0.788	1.000			
(3, 9)	0.009	0.055	0.200	0.491	0.745	1.000			
(3, 10)	0.007	0.045	0.171	0.455	0.706	1.000			
(4, 4)	0.029	0.114	0.371	0.629	0.886	0.971	1.000		
(4, 5)	0.016	0.071	0.262	0.500	0.786	0.929	0.992	1.000	
(4, 6)	0.010	0.048	0.190	0.405	0.690	0.881	0.976	1.000	
(4, 7)	0.006	0.033	0.142	0.333	0.606	0.833	0.954	1.000	
(4, 8)	0.004	0.024	0.109	0.279	0.533	0.788	0.929	1.000	
(4, 9)	0.003	0.018	0.085	0.236	0.471	0.745	0.902	1.000	
(4, 10)	0.002	0.014	0.068	0.203	0.419	0.706	0.874	1.000	
(5, 5)	0.008	0.040	0.167	0.357	0.643	0.833	0.960	0.992	1.000
(5, 6)	0.004	0.024	0.110	0.262	0.522	0.738	0.911	0.976	0.998
(5, 7)	0.003	0.015	0.076	0.197	0.424	0.652	0.854	0.955	0.992
(5, 8)	0.002	0.010	0.054	0.152	0.347	0.576	0.793	0.929	0.984
(5, 9)	0.001	0.007	0.039	0.119	0.287	0.510	0.734	0.902	0.972
(5, 10)	0.001	0.005	0.029	0.095	0.239	0.455	0.678	0.874	0.958
(6, 6)	0.002	0.013	0.067	0.175	0.392	0.608	0.825	0.933	0.987
(6, 7)	0.001	0.008	0.043	0.121	0.296	0.500	0.733	0.879	0.966
(6, 8)	0.001	0.005	0.028	0.086	0.226	0.413	0.646	0.821	0.937
(6, 9)	0.000	0.003	0.019	0.063	0.175	0.343	0.566	0.762	0.902
(6, 10)	0.000	0.002	0.013	0.047	0.137	0.288	0.497	0.706	0.864
(7, 7)	0.001	0.004	0.025	0.078	0.209	0.383	0.617	0.791	0.922
(7, 8)	0.000	0.002	0.015	0.051	0.149	0.296	0.514	0.704	0.867
(7, 9)	0.000	0.001	0.010	0.035	0.108	0.231	0.427	0.622	0.806
(7, 10)	0.000	0.001	0.006	0.024	0.080	0.182	0.355	0.549	0.743
(8, 8)	0.000	0.001	0.009	0.032	0.100	0.214	0.405	0.595	0.786
(8, 9)	0.000	0.001	0.005	0.020	0.069	0.157	0.319	0.500	0.702
(8, 10)	0.000	0.000	0.003	0.013	0.048	0.117	0.251	0.419	0.621
(9, 9)	0.000	0.000	0.003	0.012	0.044	0.109	0.238	0.399	0.601
(9, 10)	0.000	0.000	0.002	0.008	0.029	0.077	0.179	0.319	0.510
(10, 10)	0.000	0.000	0.001	0.004	0.019	0.051	0.128	0.242	0.414

Reproducida de C. Eisenhart y R. Swed, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", *Ann. Math. Stat*., **14**, 1943, con autorización del editor.

Tabla A.18 (*continuación*) $P(V \leq v^*$ cuando H_0 es verdadera) en la prueba de rachas

	v^*									
(n_1, n_2)	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
(2, 3)										
(2, 4)										
(2 5)										
(2, 6)										
(2, 7)										
(2, 8)										
(2, 9)										
(2, 10)										
(3, 3)										
(3, 4)										
(3, 5)										
(3, 6)										
(3, 7)										
(3, 8)										
(3, 9)										
(3, 10)										
(4, 4)										
(4, 5)										
(4, 6)										
(4, 7)										
(4, 8)										
(4, 9)										
(4, 10)										
(5, 5)										
(5, 6)	1.000									
(5, 7)	1.000									
(5, 8)	1.000									
(5, 9)	1.000									
(5, 10)	1.000									
(6, 6)	0.998	1.000								
(6, 7)	0.992	0.999	1.000							
(6, 8)	0.984	0.998	1.000							
(6, 9)	0.972	0.994	1.000							
(6, 10)	0.958	0.990	1.000							
(7, 7)	0.975	0.996	0.999	1.000						
(7, 8)	0.949	0.988	0.998	1.000	1.000					
(7, 9)	0.916	0.975	0.994	0.999	1.000					
(7, 10)	0.879	0.957	0.990	0.998	1.000					
(8, 8)	0.900	0.968	0.991	0.999	1.000	1.000				
(8, 9)	0.843	0.939	0.980	0.996	0.999	1.000	1.000			
(8, 10)	0.782	0.903	0.964	0.990	0.998	1.000	1.000			
(9, 9)	0.762	0.891	0.956	0.988	0.997	1.000	1.000	1.000		
(9, 10)	0.681	0.834	0.923	0.974	0.992	0.999	1.000	1.000	1.000	
(10, 10)	0.586	0.758	0.872	0.949	0.981	0.996	0.999	1.000	1.000	1.000

Tabla A.19 Tamaño muestral para límites de tolerancia no paramétricos bilaterales

	$1-\gamma$					
$1-\alpha$	**0.50**	**0.70**	**0.90**	**0.95**	**0.99**	**0.995**
0.995	336	488	777	947	1325	1483
0.99	168	244	388	473	662	740
0.95	34	49	77	93	130	146
0.90	17	24	38	46	64	72
0.85	11	16	25	30	42	47
0.80	9	12	18	22	31	34
0.75	7	10	15	18	24	27
0.70	6	8	12	14	20	22
0.60	4	6	9	10	14	16
0.50	3	5	7	8	11	12

Tabla A-25d de Wilfrid J. Dixon y Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, 1969. Reproducida con autorización de The McGraw-Hill Companies, Inc.

Tabla A.20 Tamaño muestral para límites de tolerancia no paramétricos unilaterales

	$1-\gamma$				
$1-\alpha$	**0.50**	**0.70**	**0.95**	**0.99**	**0.995**
0.995	139	241	598	919	1379
0.99	69	120	299	459	688
0.95	14	24	59	90	135
0.90	7	12	29	44	66
0.85	5	8	19	29	43
0.80	4	6	14	21	31
0.75	3	5	11	7	25
0.70	2	4	9	13	20
0.60	2	3	6	10	14
0.50	1	2	5	7	10

Tabla A-25e de Wilfrid J. Dixon y Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, 1969. Reproducida con autorización de The McGraw-Hill Companies, Inc.

Tabla A.21 Valores críticos del coeficiente de correlación de rangos de Spearman

n	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
5	0.900			
6	0.829	0.886	0.943	
7	0.714	0.786	0.893	
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.523	0.623	0.736	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478

Reproducida de E. G. Olds, "Distribution of Sums of Squares of Rank Differences for Small Samples", *Ann. Math. Stat.*, **9**, 1938, con autorización del editor.

Tabla A.22 Factores para la elaboración de gráficas de control

Observaciones en la muestra	Gráfica para promedios		Gráfica para desviaciones estándar						Gráfica para rangos				
	Factores para los límites de control		Factores para la línea central		Factores para los límites de control				Factores para la línea central		Factores para los límites de control		
n	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_3	D_4
2	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.267
3	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	2.574
4	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	2.282
5	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	2.114
6	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	2.004
7	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.076	1.924
8	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.136	1.864
9	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.184	1.816
10	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.223	1.777
11	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.256	1.744
12	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.283	1.717
13	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	0.307	1.693
14	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	0.328	1.672
15	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	0.347	1.653
16	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	0.363	1.637
17	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	0.378	1.622
18	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	0.391	1.608
19	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	0.403	1.597
20	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	0.415	1.585
21	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	0.425	1.575
22	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	0.434	1.566
23	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	0.443	1.557
24	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	0.451	1.548
25	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	0.459	4.541

Tabla A.23 La función gamma incompleta: $F(x;\alpha) = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy$

	α									
x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.6320	0.2640	0.0800	0.0190	0.0040	0.0010	0.0000	0.0000	0.0000	0.0000
2	0.8650	0.5940	0.3230	0.1430	0.0530	0.0170	0.0050	0.0010	0.0000	0.0000
3	0.9500	0.8010	0.5770	0.3530	0.1850	0.0840	0.0340	0.0120	0.0040	0.0010
4	0.9820	0.9080	0.7620	0.5670	0.3710	0.2150	0.1110	0.0510	0.0210	0.0080
5	0.9930	0.9600	0.8750	0.7350	0.5600	0.3840	0.2380	0.1330	0.0680	0.0320
6	0.9980	0.9830	0.9380	0.8490	0.7150	0.5540	0.3940	0.2560	0.1530	0.0840
7	0.9990	0.9930	0.9700	0.9180	0.8270	0.6990	0.5500	0.4010	0.2710	0.1700
8	1.0000	0.9970	0.9860	0.9580	0.9000	0.8090	0.6870	0.5470	0.4070	0.2830
9		0.9990	0.9940	0.9790	0.9450	0.8840	0.7930	0.6760	0.5440	0.4130
10		1.0000	0.9970	0.9900	0.9710	0.9330	0.8700	0.7800	0.6670	0.5420
11			0.9990	0.9950	0.9850	0.9620	0.9210	0.8570	0.7680	0.6590
12			1.0000	0.9980	0.9920	0.9800	0.9540	0.9110	0.8450	0.7580
13				0.9990	0.9960	0.9890	0.9740	0.9460	0.9000	0.8340
14				1.0000	0.9980	0.9940	0.9860	0.9680	0.9380	0.8910
15					0.9990	0.9970	0.9920	0.9820	0.9630	0.9300

A.24 Demostración de la media de la distribución hipergeométrica

Para calcular la media de la distribución hipergeométrica escribimos

$$E(X) = \sum_{x=0}^{n} x \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = k \sum_{x=1}^{n} \frac{(k-1)!}{(x-1)!(k-x)!} \cdot \frac{\binom{N-k}{n-x}}{\binom{N}{n}}$$
$$= k \sum_{x=1}^{n} \frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}.$$

Puesto que

$$\binom{N-k}{n-1-y} = \binom{(N-1)-(k-1)}{n-1-y} \quad \text{y} \quad \binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N}{n}\binom{N-1}{n-1},$$

y con $y = x - 1$, obtenemos

$$E(X) = k \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\binom{N}{n}}$$
$$= \frac{nk}{N} \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{(N-1)-(k-1)}{n-1-y}}{\binom{N-1}{n-1}} = \frac{nk}{N},$$

ya que la sumatoria representa el total de todas las probabilidades en un experimento hipergeométrico cuando

$N - 1$ artículos se seleccionan al azar de $N - 1$, de los cuales $k - 1$ se etiqueta como éxitos.

A.25 Demostración de la media y la varianza de la distribución de Poisson

Sea $\mu = \lambda t$.

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \sum_{x=1}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \mu \sum_{x=1}^{\infty} \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}.$$

Puesto que la sumatoria en el último término de la expresión anterior es la probabilidad total de una variable aleatoria de Poisson con media μ, la cual puede verse con facilidad con $y = x - 1$, es igual a 1. Por lo tanto, $E(X) = \mu$. Para calcular la varianza de X observe que

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\mu}\mu^x}{x!} = \mu^2 \sum_{x=2}^{\infty} \frac{e^{-\mu}\mu^{x-2}}{(x_2)!}.$$

Nuevamente, sea $y = x - 2$, la sumatoria en el último término de la expresión anterior es la probabilidad total de una variable aleatoria de Poisson con media μ. En consecuencia, obtenemos

$$\sigma^2 = E(X^2) - [E(X)]^2 = E[X(X-1)] + E(X) - [E(X)]^2 = \mu^2 + \mu - \mu^2 = \mu = \lambda t.$$

A.26 Prueba de la media y la varianza de la distribución gamma

Para calcular la media y la varianza de la distribución gamma comenzamos por calcular

$$E(X^k) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha+k-1} e^{-x/\beta}\, dx = \frac{\beta^{k+\alpha}\Gamma(\alpha+k)}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \frac{x^{\alpha+k-1}e^{-x/\beta}}{\beta^{k+\alpha}\Gamma(\alpha+k)}\, dx,$$

para $k = 0, 1, 2, \ldots$. Puesto que el integrando en el último término de la expresión anterior es una función de densidad gamma, con parámetros $\alpha + k$ y β, es igual a 1. Por lo tanto,

$$E(X^k) = \beta^k \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)}.$$

Si utilizamos la fórmula de recursividad de la función gamma de la página 194, obtenemos

$$\mu = \beta\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha\beta \quad \text{y} \quad \sigma^2 = E(X^2) - \mu^2 = \beta^2 \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} - \mu^2 = \beta^2\alpha(\alpha+1) - (\alpha\beta)^2 = \alpha\beta^2.$$

Apéndice B
Respuestas a los ejercicios impares (no de repaso)

Capítulo 1

1.1 *a*) Tamaño de la muestra = 15
b) Media de la muestra = 3.787
c) Mediana de la muestra = 3.6
e) $\bar{x}_{\text{tr(20)}} = 3.678$
f) Son casi iguales.

1.3 *b*) Sí, el proceso de envejecimiento redujo la resistencia a la tensión.
c) $\bar{x}_{\text{Con envejecimiento}} = 209.90$, $\bar{x}_{\text{Sin envejecimiento}} = 222.10$
d) $\bar{x}_{\text{Con envejecimiento}} = 210.00$, $\bar{x}_{\text{Sin envejecimiento}} = 221.50$
Las medias y las medianas son similares en cada grupo.

1.5 *b*) Control $\bar{x} = 5.60$, $\tilde{x} = 5.00$, $\bar{x}_{\text{tr(10)}} = 5.13$. Tratamiento: $\bar{x} = 7.60$, $\tilde{x} = 4.50$, $\bar{x}_{\text{tr(10)}} = 5.63$.
c) El valor extremo de 37 en el grupo de tratamiento desempeña un papel significativo en el cálculo de la media.

1.7 Varianza de la muestra = 0.943
Desviación estándar de la muestra = 0.971

1.9 *a*) Sin envejecimiento: varianza de la muestra = 23.66,
desviación estándar de la muestra = 4.86.
Con envejecimiento: varianza de la muestra = 42.10,
desviación estándar de la muestra = 6.49.
b) Con base en las cifras del inciso *a*), la variación en la situación "con envejecimiento" es menor que en la situación "sin envejecimiento", aunque la diferencia no sea evidente en la gráfica.

1.11 Control: varianza de la muestra = 69.38, desviación estándar de la muestra = 8.33. Tratamiento: varianza de la muestra = 128.04, desviación estándar de la muestra = 11.32.

1.13 *a*) Media = 124.3, mediana = 120
b) 175 es una observación extrema.

1.15 Sí, el valor $P = 0.03125$; probabilidad de obtener *H H H H H* con una moneda legal.

1.17 *a*) Las medias muestrales de no fumadores y fumadores son 30.32 y 43.70, respectivamente.
b) Las desviaciones estándar de la muestra de no fumadores y fumadores son 7.13 y 16.93, respectivamente.
d) Parece que a los fumadores les toma más tiempo quedarse dormidos. El tiempo que tardan los fumadores en quedarse dormidos es más variable.

1.19 *a*)

Tallo	Hojas	Frecuencia
0	22233457	8
1	023558	6
2	035	3
3	03	2
4	057	3
5	0569	4
6	0005	4

b)

Intervalo de clase	Punto medio de la clase	Frecuencia	Frecuencia relativa
0.0 – 0.9	0.45	8	0.267
1.0 – 1.9	1.45	6	0.200
2.0 – 2.9	2.45	3	0.100
3.0 – 3.9	3.45	2	0.067
4.0 – 4.9	4.45	3	0.100
5.0 – 5.9	5.45	4	0.133
6.0 – 6.9	6.45	4	0.133

c) Media muestral = 2.7967
Rango muestral = 6.3
Desviación estándar de la muestra = 2.2273

1.21 *a*) $\bar{x} = 74.02$ y $\tilde{x} = 78$
b) $s = 39.26$

1.23 *b*) $\bar{x}_{1980} = 395.10$, $\bar{x}_{1990} = 160.15$
c) Las emisiones medias cayeron entre 1980 y 1990; la variabilidad también disminuyó porque no hubo emisiones mucho más grandes.

1.25 *a*) Media muestral = 33.31
b) Mediana muestral = 26.35
d) $\bar{x}_{\text{tr}(10)} = 30.97$

Capítulo 2

2.1 *a*) $S = \{8, 16, 24, 32, 40, 48\}$
b) $S = \{-5, 1\}$
c) $S = \{T, HT, HHT, HHH\}$
d) $S = \{$África, Antártida, Asia, Australia, Europa, Norteamérica, Sudamérica$\}$
e) $S = \phi$

2.3 $A = C$

2.5 Si utilizamos un diagrama de árbol, obtenemos $S = \{1HH, 1HT, 1TH, 1TT, 2H, 2T, 3HH, 3HT, 3TH, 3TT, 4H, 4T, 5HH, 5HT, 5TH, 5TT, 6H, 6T\}$

2.7 $S_1 = \{HHHH, HHHM, HHMH, HMHH, MHHH, HHMM, HMHM, HMMH, MHMH, MMHH, MHHM, HMMM, MHMM, MMHM, MMMH, MMMM\}$;
$S_2 = \{0, 1, 2, 3, 4\}$

2.9 *a*) $A = \{1HH, 1HT, 1TH, 1TT, 2H, 2T\}$
b) $B = \{1TT, 3TT, 5TT\}$
c) $A' = \{3HH, 3HT, 3TH, 3TT, 4H, 4T, 5HH, 5HT, 5TH, 5TT, 6H, 6T\}$
d) $A' \cap B = \{3TT, 5TT\}$
e) $A \cup B = \{1HH, 1HT, 1TH, 1TT, 2H, 2T, 3TT, 5TT\}$

2.11 *a*) $S = \{H_1H_2, H_1M_1, H_1M_2, H_2H_1, H_2M_1, H_2M_2, M_1H_1, M_1H_2, M_1M_2, M_2H_1, M_2H_2, M_2M_1\}$
b) $A = \{H_1H_2, H_1M_1, H_1M_2, H_2H_1, H_2M_1, H_2M_2\}$
c) $B = \{H_1M_1, H_1M_2, H_2M_1, H_2M_2, M_1H_1, M_1H_2, M_2H_1, M_2H_2\}$
d) $C = \{M_1M_2, M_2M_1\}$
e) $A \cap B = \{H_1M_1, H_1M_2, H_2M_1, H_2M_2\}$
f) $A \cup B = \{H_1H_2, H_1M_1, H_1M_2, H_2H_1, H_2M_1, H_2M_2, M_1M_2, M_2M_1\}$

2.15 *a*) {nitrógeno, potasio, uranio, oxígeno}
b) {cobre, sodio, zinc, oxígeno}
c) {cobre, sodio, nitrógeno, potasio, uranio, zinc}
d) {cobre uranio, zinc}
e) ϕ
f) {oxígeno}

2.19 *a*) La familia experimentará fallas mecánicas, pero no recibirá una infracción por cometer una falta de tránsito, y no llegará a un lugar para acampar que esté lleno.
b) La familia recibirá una infracción por cometer una falta de tránsito y llegará a un lugar para acampar que esté lleno, pero no experimentará fallas mecánicas.
c) La familia experimentará fallas mecánicas y llegará a un lugar para acampar que esté lleno.
d) La familia recibirá una infracción por cometer una falta de tránsito, pero no llegará a un lugar para acampar que esté lleno.
e) La familia no experimentará fallas mecánicas.

2.21 18

2.23 156

2.25 20

2.27 48

2.29 210

2.31 72

2.33 *a*) 1024; *b*) 243

2.35 362,880

2.37 2880

2.39 *a*) 40,320; *b*) 336

2.41 360

2.43 24

2.45 3360

2.47 56

2.49 *a*) La suma de las probabilidades excede a 1.
b) La suma de las probabilidades es menor que 1.
c) Una probabilidad negativa.
d) La probabilidad de un corazón y de una carta negra es cero.

2.51 $S = \{\$10, \$25, \$100\}$; $P(10) = \frac{11}{20}$, $P(25) = \frac{3}{10}$, $P(100) = \frac{15}{100}$; $\frac{17}{20}$

2.53 *a*) 0.3; *b*) 0.2

2.55 10/117

2.57 *a*) 5/26; *b*) 9/26; *c*) 19/26

2.59 *a*) 94/54,145; *b*) 143/39,984

2.61 *a*) 22/25; *b*) 3/25; *c*) 17/50

2.63 *a*) 0.32; *b*) 0.68;*vc*) oficina o estudio

2.65 *a*) 0.8; *b*) 0.45; *c*) 0.55

2.67 *a*) 0.31; *b*) 0.93; *c*) 0.31

2.69 *a*) 0.009; *b*) 0.999; *c*) 0.01

2.71 *a*) 0.048; *b*) \$50,000; *c*) \$12,500

2.73 *a*) La probabilidad de que un convicto que vende drogas también cometa un robo a mano armada.
b) La probabilidad de que un convicto que cometa un robo a mano armada no venda drogas.
c) La probabilidad de que un convicto que no vende drogas tampoco cometa un robo a mano armada.

2.75 *a*) 14/39; *b*) 95/112

2.77 *a*) 5/34; *b*) 3/8

2.79 *a*) 0.018; *b*) 0.614; *c*) 0.166; *d*) 0.479

2.81 *a*) 0.35; *b*) 0.875; *c*) 0.55

2.83 *a*) 9/28; *b*) 3/4; *c*) 0.91

2.85 0.27

2.87 5/8

2.89 *a*) 0.0016; *b*) 0.9984

2.91 *a*) 91/323; *b*) 91/323

2.93 *a*) 0.75112; *b*) 0.2045

2.95 0.0960

2.97 0.40625

2.99 0.1124

2.101 0.857

Capítulo 3

3.1 Discreta; continua; continua; discreta; discreta; continua.

3.3

Espacio muestral	w
HHH	3
HHT	1
HTH	1
THH	1
HTT	−1
THT	−1
TTH	−1
TTT	−3

3.5 *a*) 1/30; *b*) 1/10

3.7 *a*) 0.68; *b*) 0.375

3.9 *b*) 19/80

3.11

x	0	1	2
$f(x)$	$\frac{2}{7}$	$\frac{4}{7}$	$\frac{1}{7}$

3.13
$$F(x) = \begin{cases} 0, & \text{para } x < 0, \\ 0.41, & \text{para } 0 \le x < 1, \\ 0.78, & \text{para } 1 \le x < 2, \\ 0.94, & \text{para } 2 \le x < 3, \\ 0.99, & \text{para } 3 \le x < 4, \\ 1, & \text{para } x \ge 4 \end{cases}$$

3.15
$$F(x) = \begin{cases} 0, & \text{para } x < 0, \\ \frac{2}{7}, & \text{para } 0 \le x < 1, \\ \frac{6}{7}, & \text{para } 1 \le x < 2, \\ 1, & \text{para } x \ge 2 \end{cases}$$

a) 4/7; *b*) 5/7

3.17 *b*) 1/4; *c*) 0.3

3.19
$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{x-1}{2}, & 1 \le x < 3; \ 1/4 \\ 1, & x \ge 3 \end{cases}$$

3.21 *a*) $3/2$; *b*) $F(x) = \begin{cases} 0, & x < 0 \\ x^{3/2}, & 0 \le x < 1; \\ 1, & x \ge 1 \end{cases}$ 0.3004

3.23 $F(w) = \begin{cases} 0, & \text{para } w < -3, \\ \frac{1}{27}, & \text{para } -3 \le w < -1, \\ \frac{7}{27}, & \text{para } -1 \le w < 1, \\ \frac{19}{27}, & \text{para } 1 \le w < 3, \\ 1, & \text{para } w \ge 3 \end{cases}$

a) $20/27$; *b*) $2/3$

3.25

t	20	25	30
$P(T = t)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

3.27 *a*) $F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp(-x/2000), & x \ge 0 \end{cases}$

b) 0.6065; *c*) 0.6321

3.29 *b*) $F(x) = \begin{cases} 0, & x < 1, \\ 1 - x^{-3}, & x \ge 1 \end{cases}$

c) 0.0156

3.31 *a*) 0.2231; *b*) 0.2212

3.33 *a*) $k = 280$; *b*) 0.3633; *c*) 0.0563

3.35 *a*) 0.1528; *b*) 0.0446

3.37 *a*) $1/36$; *b*) $1/15$

3.39 *a*)

	$f(x, y)$	x = 0	1	2	3
y	0	0	$\frac{3}{70}$	$\frac{9}{70}$	$\frac{3}{70}$
	1	$\frac{2}{70}$	$\frac{18}{70}$	$\frac{18}{70}$	$\frac{2}{70}$
	2	$\frac{3}{70}$	$\frac{9}{70}$	$\frac{3}{70}$	0

b) $1/2$

3.41 *a*) $1/16$; *b*) $g(x) = 12x(1-x)^2$, para $0 \le x \le 1$; *c*) $1/4$

3.43 *a*) $3/64$; *b*) $1/2$

3.45 0.6534

3.47 *a*) Dependiente; *b*) $1/3$

3.49 *a*)

x	1	2	3
$g(x)$	0.10	0.35	0.55

b)

y	1	2	3
$h(y)$	0.20	0.50	0.30

c) 0.2857

3.51 *a*)

	$f(x, y)$	x = 0	1	2	3
y	0	$\frac{1}{55}$	$\frac{6}{55}$	$\frac{6}{55}$	$\frac{1}{55}$
	1	$\frac{6}{55}$	$\frac{16}{55}$	$\frac{6}{55}$	0
	2	$\frac{6}{55}$	$\frac{6}{55}$	0	0
	3	$\frac{1}{55}$	0	0	0

b) $42/55$

3.53 $5/8$

3.55 Independiente

3.57 *a*) 3; *b*) $21/512$

3.59 Dependiente

Capítulo 4

4.1 0.88

4.3 25 centavos

4.5 \$1.23

4.7 \$500

4.9 \$6900

4.11 $(\ln 4)/\pi$

4.13 100 horas

4.15 0

4.17 209

4.19 \$1855

4.21 \$833.33

4.23 *a*) 35.2; *b*) $\mu_X = 3.20$; $\mu_Y = 3.00$

4.25 2

4.27 2000 horas

4.29 *b*) $3/2$

4.31 *a*) $1/6$; *b*) $(5/6)^5$

4.33 \$5,250,000

4.35 0.74

4.37 $1/18$; en términos de utilidad real la varianza es $\frac{1}{18}(5000)^2$

4.39 $1/6$

4.41 118.9

4.43 $\mu_Y = 10$; $\sigma_Y^2 = 144$

4.45 0.01

4.47 -0.0062

4.49 $\sigma_X^2 = 0.8456$, $\sigma_X = 0.9196$

4.51 $-1/\sqrt{5}$

4.53 $\mu_{g(X)} = 10.33$, $\sigma_{g(X)} = 6.66$

4.55 \$0.80

4.57 209

4.59 $\mu = 7/2$, $\sigma^2 = 15/4$

4.61 $3/14$

4.63 52

4.65 *a*) 7; *b*) 0; *c*) 12.25

4.67 $46/63$

4.69 *a*) $E(X) = E(Y) = 1/3$ y $\text{Var}(X) = \text{Var}(Y) = 4/9$; *b*) $E(Z) = 2/3$ y $\text{Var}(Z) = 8/9$

4.71 *a*) 4; *b*) 32; 16

4.73 Mediante cálculo directo, $E(e^Y) = 1884.32$. Si usamos la aproximación de ajuste de segundo orden, $E(e^Y) \approx 1883.38$, que se acerca mucho al valor real.

4.75 0.03125

4.77 *a*) A lo sumo $4/9$; *b*) al menos $5/9$; *c*) al menos $21/25$; *d*) 10

Capítulo 5

5.1 $\mu = \frac{1}{k}\sum_{i=1}^{k} x_i$, $\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2$

5.3 $f(x) = \frac{1}{10}$ para $x = 1,2,...,10$ y $f(x) = 0$ en otro caso; $3/10$

5.5 *a*) 0.0480; *b*) 0.2375; *c*) $P(X = 5 \mid p = 0.3) = 0.1789$, $P = 0.3$ es razonable.

5.7 *a*) 0.0474; *b*) 0.0171

5.9 *a*) 0.7073; *b*) 0.4613; *c*) 0.1484

5.11 0.1240

5.13 0.8369

5.15 *a*) 0.0778; *b*) 0.3370; *c*) 0.0870

5.17 $\mu = 3.5$, $\sigma^2 = 1.05$

5.19 $f(x_1, x_2, x_3) = \binom{n}{x_1, x_2, x_3} 0.35^{x_1}\, 0.05^{x_2}\, 0.60^{x_3}$

5.21 0.0095

5.23 0.0077

5.25 0.8670

5.27 *a*) 0.2852; *b*) 0.9887; *c*) 0.6083

5.29 $5/14$

5.31 $h(x; 6, 3, 4) = \dfrac{\binom{4}{x}\binom{2}{3-x}}{\binom{6}{3}}$, para $x = 1, 2, 3$; $P(2 \le X \le 3) = 4/5$

5.33 *a*) 0.3246; *b*) 0.4496

5.35 0.9517

5.37 *a*) 0.6815; *b*) 0.1153

5.39 0.9453

5.41 0.6077

5.43 *a*) $4/33$; *b*) $8/165$

5.45 0.2315

5.47 *a*) 0.3991; *b*) 0.1316

5.49 0.0515

5.51 $63/64$

5.53 *a*) 0.3840; *b*) 0.0067

5.55 *a*) 0.0630; *b*) 0.9730

5.57 *a*) 0.1429; *b*) 0.1353

5.59 *a*) 0.1638; *b*) 0.032

5.61 0.2657

5.63 $\mu = 6$, $\sigma^2 = 6$

5.65 *a*) 0.2650; *b*) 0.9596

5.67 *a*) 0.8243; *b*) 14

5.69 4

5.71 5.53×10^{-4}; $\mu = 7.5$

5.73 *a*) 0.0137; *b*) 0.0830

5.75 0.4686

Capítulo 6

6.3 *a*) 0.6; *b*) 0.7; *c*) 0.5

6.5 *a*) 0.0823; *b*) 0.0250; *c*) 0.2424; *d*) 0.9236; *e*) 0.8133; *f*) 0.6435

6.7 *a*) 0.54; *b*) –1.72; *c*) 1.28

6.9 *a*) 0.1151; *b*) 16.1; *c*) 20.275; *d*) 0.5403

6.11 *a*) 0.0548; *b*) 0.4514; *c*) 23 tazas *d*) 189.95 mililitros

6.13 *a*) 0.8980; *b*) 0.0287; *c*) 0.6080

6.15 *a*) 0.0571; *b*) 99.11%; *c*) 0.3974; *d*) 27.952 minutos; *e*) 0.0092

6.17 6.24 años

6.19 *a*) 51%; *b*) \$18.37

6.21 *a*) 0.0401; *b*) 0.0244

6.23 26 estudiantes

6.25 *a*) 0.3085; *b*) 0.0197

6.27 *a*) 0.9514; *b*) 0.0668

6.29 *a*) 0.1171; *b*) 0.2049

6.31 0.1357

6.33 *a*) 0.0778; *b*) 0.0571; *c*) 0.6811

6.35 *a*) 0.8749; *b*) 0.0059

6.37 *a*) 0.0228; *b*) 0.3974

6.41 $2.8e^{-1.8} - 3.4e^{-2.4} = 0.1545$

6.43 *a*) $\mu = 6$; $\sigma^2 = 18$; *b*) entre 0 y 14.485 millones de litros

6.45 $\sum_{x=4}^{6} \binom{6}{x} (1 - e^{-3/4})^x (e^{-3/4})^{6-x} = 0.3968$

6.47 *a*) $\sqrt{\pi/2} = 1.2533$ años; *b*) e^{-2}

6.49 *a*) Media = 0.25, mediana = 0.206; *b*) varianza = 0.0375; *c*) 0.2963

6.51 $e^{-4} = 0.0183$

6.53 *a*) $\mu = \alpha\beta = 50$; *b*) $\sigma^2 = \alpha\beta^2 = 500$; $\sigma = \sqrt{500}$ *c*) 0.815

6.55 *a*) 0.1889; *b*) 0.0357

6.57 Media $= e^6$, varianza $= e^{12}(e^4 - 1)$

6.59 *a*) e^{-5}; *b*) $\beta = 0.2$

Capítulo 7

7.1 $g(y) = 1/3$, para $y = 1, 3, 5$

7.3 $$g(y_1, y_2) = \binom{2}{\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}, 2-y_1} \times \left(\frac{1}{4}\right)^{(y_1+y_2)/2} \left(\frac{1}{3}\right)^{(y_1-y_2)/2} \left(\frac{5}{12}\right)^{2-y_1};$$ para $y_1 = 0, 1, 2$; $y_2 = -2, -1, 0, 1, 2$; $y_2 \le y_1$; $y_1 + y_2 = 0, 2, 4$

7.7 Distribución gamma con $\alpha = 3/2$ y $\beta = m/2b$

7.9 *a*) $g(y) = 32/y^3$, para $y > 4$; *b*) 1/4

7.11 $h(z) = 2(1 - z)$, para $0 < z < 1$

7.13 $h(w) = 6 + 6w - 12w^{1/2}$, para $0 < w < 1$

7.15 $$g(y) = \begin{cases} \frac{2}{9\sqrt{y}}, & 0 < y < 1, \\ \frac{\sqrt{y}+1}{9\sqrt{y}}, & 1 < y < 4 \end{cases}$$

7.19 Ambas son iguales a μ.

7.23 *a*) Gamma(2, 1); *b*) Uniforme(0, 1)

Capítulo 8

8.1 *a*) Las respuestas de todas las personas en Richmond que tienen teléfono;
b) Resultados para un número grande o infinito de lanzamientos de una moneda;
c) Periodo de vida de tal calzado deportivo cuando es utilizado en el torneo profesional;
d) Todos los posibles intervalos de tiempo para esta abogada que maneja desde su casa hasta su oficina.

8.3 *a*) $\bar{x} = 3.2$ segundos; *b*) $\tilde{x} = 3.1$ segundos

8.5 *a*) $\bar{x} = 2.4$; *b*) $\tilde{x} = 2$; *c*) $m = 3$

8.7 *a*) 53.75; *b*) 75 y 100

8.9 *a*) El rango es 10; *b*) $s = 3.307$

8.11 *a*) 2.971; *b*) 2.971

8.13 $s = 0.585$

8.15 *a*) 45.9; *b*) 5.1

8.17 0.3159

8.19 *a*) La varianza se reduce de 0.49 a 0.16;
b) La varianza se incrementa de 0.04 a 0.64.

8.21 Sí.

8.23 *a*) $\mu = 5.3$; $\sigma^2 = 0.81$;
b) $\mu_{\bar{X}} = 5.3$; $\sigma^2_{\bar{X}} = 0.0225$;
c) 0.9082

8.25 *a*) 0.6898; *b*) 7.35

8.29 0.5596

8.31 *a*) La probabilidad de que el tiempo promedio de secado sea mayor que 1.0 es 0.0013; *b*) 13

8.33 *a*) $1/2$; *b*) 0.3085

8.35 $P(\bar{X} \leq 775 \,|\, \mu = 760) = 0.9332$

8.37 *a*) 27.488; *b*) 18.475; *c*) 36.415

8.39 *a*) 0.297; *b*) 32.852; *c*) 46.928

8.41 *a*) 0.05; *b*) 0.94

8.45 *a*) 0.975; *b*) 0.10; *c*) 0.875; *d*) 0.99

8.47 *a*) 2.500; *b*) 1.319; *c*) 1.714

8.49 No; $\mu > 20$

8.51 *a*) 2.71; *b*) 3.51; *c*) 2.92; *d*) 0.47; *e*) 0.34

8.53 La razón F es 1.44. Las varianzas no son significativamente diferentes.

Capítulo 9

9.1 56

9.3 $0.3097 < \mu < 0.3103$

9.5 *a*) $22{,}496 < \mu < 24{,}504$; *b*) error ≤ 1004

9.7 35

9.9 $10.15 < \mu < 12.45$

9.11 $0.978 < \mu < 1.033$

9.13 $47.722 < \mu < 49.278$

9.15 (13,075, 33,925)

9.17 (6.05, 16.55)

9.19 323.946 a 326.154

9.21 Límite superior de predicción: 9.42;
límite superior de tolerancia: 11.72

9.25 Sí, el valor de 6.9 está fuera del intervalo de predicción.

9.27 *a*) (0.9876, 1.0174);
b) (0.9411, 1.0639);
c) (0.9334, 1.0716)

9.35 $2.9 < \mu_1 - \mu_2 < 7.1$

9.37 $2.80 < \mu_1 - \mu_2 < 3.40$

9.39 $1.5 < \mu_1 - \mu_2 < 12.5$

9.41 $0.70 < \mu_1 - \mu_2 < 3.30$

9.43 $-6536 < \mu_1 - \mu_2 < 2936$

9.45 $(-0.74, 6.30)$

9.47 $(-6.92, 36.70)$

9.49 $0.54652 < \mu_B - \mu_A < 1.69348$

9.51 Método 1: $0.194 < p < 0.262$; método 2: $0.1957 < p < 0.2639$

9.53 *a*) $0.498 < p < 0.642$; *b*) error ≤ 0.072

9.55 *a*) $0.739 < p < 0.961$; *b*) no

9.57 *a*) $0.644 < p < 0.690$; *b*) error ≤ 0.023

9.59 2576

9.61 160

9.63 9604

9.65 $-0.0136 < p_F - p_M < 0.0636$

9.67 $0.0011 < p_1 - p_2 < 0.0869$

9.69 $(-0.0849, 0.0013)$; no es significativamente diferente.

9.71 $0.293 < \sigma^2 < 6.736$; la afirmación es válida

9.73 $3.472 < \sigma^2 < 12.804$

9.75 $9.27 < \sigma < 34.16$

9.77 $0.549 < \sigma_1 / \sigma_2 < 2.690$

9.79 $0.016 < \sigma_1^2/\sigma_2^2 < 0.454$; no

9.81 $\frac{1}{n}\sum_{i=1}^{n} x_i$

9.83 $\hat{\beta} = \bar{x}/5$

9.85 $\hat{\theta} = \text{máx}\{x_1, \ldots, x_n\}$

9.87 $x \ln p + (1 - x) \ln(1 - p)$. Sea la derivada con respecto a $p = 0$; $\hat{p} = x = 1.0$

Capítulo 10

10.1 *a*) Concluya que menos de 30% del público es alérgico a ciertos productos de queso cuando, de hecho, 30% o más es alérgico.
b) Concluya que al menos 30% del público es alérgico a ciertos productos de queso cuando, de hecho, menos de 30% es alérgico.

10.3 *a*) La empresa no es culpable;
b) la empresa es culpable.

10.5 *a*) 0.0559;
b) $\beta = 0.0017$; $\beta = 0.00968$; $\beta = 0.5557$

10.7 *a*) 0.1286;
b) $\beta = 0.0901$; $\beta = 0.0708$.
c) La probabilidad de un error tipo I es algo grande.

10.9 *a*) $\alpha = 0.0850$; *b*) $\beta = 0.3410$

10.11 *a*) $\alpha = 0.1357$; *b*) $\beta = 0.2578$

10.13 $\alpha = 0.0094$; $\beta = 0.0122$

10.15 *a*) $\alpha = 0.0718$; *b*) $\beta = 0.1151$

10.17 *a*) $\alpha = 0.0384$; *b*) $\beta = 0.5$; $\beta = 0.2776$

10.19 $z = -2.76$; sí, $\mu < 40$ meses; valor $P = 0.0029$

10.21 $z = -1.64$; valor $P = 0.10$

10.23 $t = 0.77$; no rechace H_0.

10.25 $z = 8.97$; sí, $\mu > 20{,}000$ kilómetros; valor $P < 0.001$

10.27 $t = 12.72$; valor $P < 0.0005$; rechace H_0.

10.29 $t = -1.98$; valor $P = 0.0312$; rechace H_0.

10.31 $z = -2.60$; concluya que $\mu_A - \mu_B \leq 12$ kilogramos.

10.33 $t = 1.50$; no hay evidencia suficiente para concluir que el incremento en la concentración de sustrato causaría un incremento en la velocidad media de más de 0.5 micromoles por 30 minutos.

10.35 $t = 0.70$; no hay suficiente evidencia que apoye la conclusión de que el suero es efectivo.

10.37 $t = 2.55$; rechace H_0: $\mu_1 - \mu_2 > 4$ kilómetros.

10.39 $t' = 0.22$; no rechace H_0.

10.41 $t' = 2.76$; rechace H_0.

10.43 $t = -2.53$; rechace H_0; la afirmación es válida.

10.45 $t = 2.48$; valor $P < 0.02$; rechace H_0.

10.47 $n = 6$

10.49 $78.28 \approx 79$

10.51 5

10.53 *a*) H_0: $M_{\text{caliente}} - M_{\text{frío}} = 0$,
H_1: $M_{\text{caliente}} - M_{\text{frío}} \neq 0$;
b) t apareada, $t = 0.99$; valor $P > 0.30$; no rechace H_0.

10.55 valor $P = 0.4044$ = (con una prueba de una cola); no se refuta la afirmación.

10.57 $z = 1.44$; no rechace H_0.

10.59 $z = -5.06$ con valor $P \approx 0$; concluya que menos de una quinta parte de los hogares se calienta con petróleo.

10.61 $z = 0.93$ con valor $P = P(Z > 0.93) = 0.1762$; no hay evidencia suficiente para concluir que la nueva medicina es eficaz.

10.63 $z = 2.36$ con valor $P = 0.0182$; sí, la diferencia es significativa.

10.65 $z = 1.10$ con valor $P = 0.1357$; no tenemos evidencia suficiente para concluir que el cáncer de mama es más frecuente en las comunidades urbanas.

10.67 $\chi^2 = 18.13$ con valor $P = 0.0676$ (de los resultados por computadora); no rechace H_0: $\sigma^2 = 0.03$.

10.69 $\chi^2 = 63.75$ con valor $P = 0.8998$ (de los resultados por computadora); no rechace H_0.

10.71 $\chi^2 = 42.37$ con valor $P = 0.0117$ (de los resultados por computadora); la máquina está fuera de control.

10.73 $f = 1.33$ con valor $P = 0.3095$ (de los resultados por computadora); no rechace H_0: $\sigma_1 = \sigma_2$.

10.75 $f = 0.086$ con valor $P = 0.0328$ (de los resultados por computadora); rechace H_0: $\sigma_1 = \sigma_2$ a un nivel mayor que 0.0328.

10.77 $f = 19.67$ con valor $P = 0.0008$ (de los resultados por computadora); rechace H_0: $\sigma_1 = \sigma_2$.

10.79 $\chi^2 = 10.14$; rechace H_0, la razón no es 5:2:2:1.

10.81 $\chi^2 = 4.47$; no hay evidencia suficiente para afirmar que el dado esté cargado.

10.83 $\chi^2 = 3.125$; no rechace H_0: distribución geométrica.

10.85 $\chi^2 = 5.19$; no rechace H_0: distribución normal.

10.87 $\chi^2 = 5.47$; no rechace H_0.

10.89 $\chi^2 = 124.59$; sí, la ocurrencia de estos tipos de delitos depende del distrito de la ciudad.

10.91 $\chi^2 = 5.92$ con valor $P = 0.4332$; no rechace H_0.

10.93 $\chi^2 = 31.17$ con valor $P < 0.0001$; las actitudes no son homogéneas.

10.95 $\chi^2 = 1.84$; no rechace H_0.

Capítulo 11

11.1 *a*) $b_0 = 64.529$, $b_1 = 0.561$;
b) $\hat{y} = 81.4$

11.3 *a*) $\hat{y} = 5.8254 + 0.5676x$;
c) $\hat{y} = 34.205$ a 50°C

11.5 *a*) $\hat{y} = 6.4136 + 1.8091x$;
b) $\hat{y} = 9.580$ a temperatura 1.75

11.7 *b*) $\hat{y} = 31.709 + 0.353x$

11.9 *b*) $\hat{y} = 343.706 + 3.221x$;
c) $\hat{y} = \$456$ con costos de publicidad $= \$35$

11.11 *b*) $\hat{y} = -1847.633 + 3.653x$

11.13 *a*) $\hat{y} = 153.175 - 6.324x$;
b) $\hat{y} = 123$ para $x = 4.8$ unidades

11.15 *a*) $s^2 = 176.4$;
b) $t = 2.04$; no rechace H_0: $\beta_1 = 0$

11.17 *a*) $s^2 = 0.40$;
b) $4.324 < \beta_0 < 8.503$;
c) $0.446 < \beta_1 < 3.172$

11.19 *a*) $s^2 = 6.626$;
b) $2.684 < \beta_0 < 8.968$;
c) $0.498 < \beta_1 < 0.637$

11.21 $t = -2.24$; rechace H_0 y concluya $\beta < 6$

11.23 *a*) $24.438 < \mu_{Y|24.5} < 27.106$;
b) $21.88 < y_0 < 29.66$

11.25 $7.81 < \mu_{Y|1.6} < 10.81$

11.27 *a*) 17.1812 mpg;
b) no, el intervalo de confianza de 95% sobre la media mpg es (27.95, 29.60);
c) las millas por galón probablemente excederán a 18.

11.29 *b*) $\hat{y} = 3.4156x$

11.31 El valor f para probar la falta de ajuste es 1.58 y se concluye que no se rechaza H_0. Por lo tanto, la prueba de falta de ajuste es insignificante.

11.33 *a*) $\hat{y} = 2.003x$;
b) $t = 1.40$, no rechace H_0.

11.35 $f = 1.71$ y valor $P = 0.2517$; la regresión es lineal.

11.37 *a*) $b_0 = 10.812$, $b_1 = -0.3437$;
b) $f = 0.43$; la regresión es lineal.

11.39 *a*) $\hat{P} = -11.3251 - 0.0449T$;
b) sí;
c) $R^2 = 0.9355$;
d) sí

11.41 *b*) $\hat{N} = -175.9025 + 0.0902Y$; $R^2 = 0.3322$

11.43 $r = 0.240$

11.45 *a*) $r = -0.979$;
b) Valor $P = 0.0530$; no rechace H_0 a un nivel de 0.025;
c) 95.8%

11.47 *a*) $r = 0.784$;
b) rechace H_0 y concluya que $\rho > 0$;
c) 61.5%.

Capítulo 12

12.1 $\hat{y} = 0.5800 + 2.7122x_1 + 2.0497x_2$

12.3 *a*) $\hat{y} = 27.547 + 0.922x_1 + 0.284x_2$;
b) $\hat{y} = 84$ para $x_1 = 64$ y $x_2 = 4$

12.5 *a*) $\hat{y} = -102.7132 + 0.6054x_1 + 8.9236x_2 + 1.4374x_3 + 0.0136x_4$;
b) $\hat{y} = 287.6$

12.7 $\hat{y} = 141.6118 - 0.2819x + 0.0003x^2$

12.9 *a*) $\hat{y} = 56.4633 + 0.1525x - 0.00008x^2$;
b) $\hat{y} = 86.7\%$ cuando la temperatura es de 225°C

12.11 $\hat{y} = -6.5122 + 1.9994x_1 - 3.6751x_2 + 2.5245x_3 + 5.1581x_4 + 14.4012x_5$

12.13 *a*) $\hat{y} = 350.9943 - 1.2720x_1 - 0.1539x_2$;
b) $\hat{y} = 140.9$

12.15 $\hat{y} = 3.3205 + 0.4210x_1 - 0.2958x_2 + 0.0164x_3 + 0.1247x_4$.

12.17 0.1651

12.19 242.72

12.21 *a*) $\hat{\sigma}^2_{B_2} = 28.0955$; *b*) $\hat{\sigma}_{B_1\,B_2} = -0.0096$

12.23 $t = 5.91$ con valor $P = 0.0002$. Rechace H_0 y asevere que $\beta_1 \neq 0$.

12.25 $0.4516 < \mu_{Y\,|\,x_1 = 900,\,x_2 = 1} < 1.2083$ y $-0.1640 < y_0 < 1.8239$

12.27 $263.7879 < \mu_{Y\,|x_1 = 75,\,x_2 = 24,\,x_3 = 90,\,x_4 = 98} < 311.3357$ y $243.7175 < y_0 < 331.4062$

12.29 *a*) $t = -1.09$ con valor $P = 0.3562$;
b) $t = -1.72$ con valor $P = 0.1841$;
c) sí; no hay suficiente evidencia que demuestre que los valores de x_1 y x_2 son significativos.

12.31 $R^2 = 0.9997$

12.33 $f = 5.106$ con valor $P = 0.0303$; la regresión no es significativa en el nivel 0.01.

12.35 $f = 34.90$ con valor $P = 0.0002$; rechace H_0 y concluya que $\beta_1 > 0$.

12.37 $f = 10.18$ con valor $P < 0.01$; x_1 y x_2 son significativos en la presencia de x_3 y x_4.

12.39 El modelo de dos variables es mejor.

12.41 Primer modelo: $R^2_{\text{adj}} = 92.7\%$, C.V. = 9.0385. Segundo modelo: $R^2_{\text{adj}} = 98.1\%$, C.V. = 4.6287. La prueba F parcial revela un valor $P = 0.0002$; el modelo 2 es mejor.

12.43 No hay mucha diferencia entre utilizar x_2 solo y usar x_1 y x_2 juntos, ya que R^2_{adj} constituye 0.7696 en comparación con 0.7591, respectivamente.

12.45 *a*) $\widehat{\text{mpg}} = 5.9593 - 0.00003773$ odómetro $+ 0.3374$ octanaje $- 12.6266z_1 - 12.9846z_2$;
b) sedán
c) no son significativamente diferentes.

12.47 *b*) $\hat{y} = 4.690$ segundos;
c) $4.450 < \mu_{Y|\{180,260\}} < 4.930$

12.49 $\hat{y} = 2.1833 + 0.9576x_2 + 3.3253x_3$

12.51 *a*) $\hat{y} = -587.211 + 428.433x$;
b) $\hat{y} = 1180 - 191.691x + 35.20945x^2$;
c) modelo cuadrático

12.53 $\hat{\sigma}^2_{B_1} = 20{,}588$; $\hat{\sigma}^2_{B_{11}} = 62.6502$; $\hat{\sigma}_{B_1,B_{11}} = -1103.5$

12.55 *a*) Es mejor el modelo de intersección.

12.57 *a*) $\hat{y} = 3.1368 + 0.6444x_1 - 0.0104x_2 + 0.5046x_3 - 0.1197x_4 - 2.4618x_5 + 1.5044x_6$;
b) $\hat{y} = 4.6563 + 0.5133x_3 - 0.1242x_4$;
c) Criterio C_p: variables x_1 y x_2 con $s^2 = 0.7317$ y $R^2 = 0.6476$; criterio s^2: variables x_1, x_3 y x_4 con $s^2 = 0.7251$ y $R^2 = 0.6726$:
d) $\hat{y} = 4.6563 + 0.5133x_3 - 0.1242x_4$; éste no pierde mucho en s^2 y R^2.
e) dos observaciones tienen valores grandes de R de Student y deben verificarse.

12.59 *a*) $\hat{y} = 125.8655 + 7.7586x_1 + 0.0943x_2 - 0.0092x_1x_2$;
b) el modelo que sólo contiene x_2 es el mejor.

12.61 *a*) $\hat{p} = (1 + e^{2.9949 - 0.0308\,x})^{-1}$;
b) 1.8515

Capítulo 13

13.1 $f = 0.31$; no hay evidencia suficiente para apoyar la hipótesis de que existen diferencias entre las 6 máquinas.

13.3 $f = 14.52$; sí, la diferencia es significativa.

13.5 $f = 8.38$; las actividades específicas promedio difieren de manera significativa.

13.7 $f = 2.25$; no hay evidencia suficiente para apoyar la hipótesis de que las diferentes concentraciones de $MgNH_4PO_4$ influyen significativamente en la altura que alcanzan los crisantemos.

13.9 $b = 0.79 > b_4(0.01, 4, 4, 4, 9) = 0.4939$. No rechace H_0. No hay suficiente evidencia para afirmar que las varianzas son diferentes.

13.11 $b = 0.7822 < b_4(0.05, 9, 8, 15) = 0.8055$. Las varianzas son significativamente diferentes.

13.13 *a*) Valor $P < 0.0001$, significativa,
b) para el contraste 1 contra 2, valor $P < 0.0001Z$, significativamente diferentes; para el contraste 3 contra 4, valor $P = 0.0648$, no es significativamente diferente.

13.15 A continuación se presentan los resultados para la prueba de Tukey

$\bar{y}_{4.}$	$\bar{y}_{3.}$	$\bar{y}_{1.}$	$\bar{y}_{5.}$	$\bar{y}_{2.}$
2.98	4.30	5.44	6.96	7.90

13.17 *a*) valor $P = 0.0121$; sí, hay una diferencia significativa;

Disminución	De Hess modificado	Remoción del sustrato de Kicknet	Surber	Kicknet

13.19 $f = 70.27$ con valor $P < 0.0001$; rechace H_0.

$\bar{x}_0$	$\bar{x}_{25}$	$\bar{x}_{100}$	$\bar{x}_{75}$	$\bar{x}_{50}$
55.167	60.167	64.167	70.500	72.833

La temperatura es importante; tanto 75° como 50°(C) producen baterías con vida activa significativamente más larga.

13.21 La absorción media para el agregado 4 es significativamente menor que para el otro agregado.

13.23 Al comparar el control con 1 y 2, significativo; al comparar el control con 3 y 4: insignificante

13.25 f(fertilizante) = 6.11; existe una diferencia significativa entre los fertilizantes

13.27 $f = 5.99$; el porcentaje de aditivos extranjeros no es el mismo para las tres marcas de mermelada; marca *A*.

13.29 Valor $P < 0.0001$; significativo

13.31 Valor $P = 0.0023$; significativo

13.33 Valor $P = 0.1250$; no significativo

13.35 Valor $P < 0.0001$;
$f = 122.37$; la cantidad de tinta sí influye en el color de la tela.

13.37 *a*) $y_{ij} = \mu + A_i + \epsilon_{ij}$, $A_i \sim n(x; 0, \sigma_\alpha)$, $\epsilon_{ij} \sim n(x; 0, \sigma)$;
b) $\hat{\sigma}_\alpha^2 = 0$ (el componente de la varianza estimada es -0.00027; $\hat{\sigma}^2 = 0.0206$.

13.39 *a*) $f = 14.9$; los operadores difieren significativamente;
b) $\hat{\sigma}_\alpha^2 = 28.91$; $s^2 = 8.32$.

13.41 *a*) $y_{ij} = \mu + A_i + \epsilon_{ij}$, $A_i \sim n(x; 0, \sigma_\alpha)$;
b) sí, $f = 5.63$ con un valor $P = 0.0121$;
c) hay un componente significativo de varianza del telar.

Capítulo 14

14.1 *a*) $f = 8.13$; significativo;
b) $f = 5.18$; significativo;
c) $f = 1.63$; no significativo

14.3 *a*) $f = 14.81$; significativo;
b) $f = 9.04$; significativo;
c) $f = 0.61$; no significativo;

14.5 *a*) $f = 34.40$; significativo;
b) $f = 26.95$; significativo;
c) $f = 20.30$; significativo;

14.7 Prueba del efecto de la temperatura: $f_1 = 10.85$ con valor $P = 0.0002$;
Prueba del efecto de la cantidad de catalizador: $f_2 = 46.63$ con valor $P < 0.0001$;
Prueba del efecto de interacción: $f = 2.06$ con valor $P = 0.074$.

14. 9 *a*)

Fuente de variación	**gl**	**Suma de cuadrados**	**Cuadrados medios**	***f***	***P***
Velocidad de corte	1	12.000	12.000	1.32	0.2836
Geometría de la herramienta	1	675.000	675.000	74.31	< 0.0001
Interacción	1	192.000	192.000	21.14	0.0018
Error	8	72.667	9.083		
Total	11	951.667			

b) El efecto de la interacción oculta el efecto de la velocidad de corte;
c) $f_{\text{geometría de la herramienta}=1} = 16.51$ y valor $P = 0.0036$;
d) $f_{\text{geometría de la herramienta}=2} = 5.94$ y valor $P = 0.0407$.

14.11 *a*)

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	*f*	*P*
Método	1	0.000104	0.000104	6.57	0.0226
Laboratorio	6	0.008058	0.001343	84.70	< 0.0001
Interacción	6	0.000198	0.000033	2.08	0.1215
Error	14	0.000222	0.000016		
Total	27	0.008582			

b) La interacción no es significativa;
c) Ambos efectos principales son significativos;
e) $f_{\text{laboratorio}=1} = 0.01576$ y valor $P = 0.9019$; no hay diferencia significativa entre los métodos en el laboratorio 1;
$f_{\text{geometría de la herramienta}=2} = 9.081$ y valor $P = 0.0093$.

14.13 *b*)

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	*f*	*P*
Tiempo	1	0.060208	0.060208	157.07	< 0.0001
Tratamiento	1	0.060208	0.060208	157.07	< 0.0001
Interacción	1	0.000008	0.000008	.02	0.8864
Error	8	0.003067	0.000383		
Total	11	0.123492			

c) Tanto el tiempo como el tratamiento influyen significativamente en la absorción del magnesio, aunque no existe interacción significativa entre ambos.
d) $Y = \mu + \beta_T$ Tiempo $+ \beta_Z Z + \beta_{TZ}$ Tiempo $Z + \epsilon$, donde $Z = 1$ cuando el tratamiento $= 1$ y $Z = 0$ cuando el tratamiento $= 2$;
e) $f = 0.02$ con valor $P = 0.8864$; la interacción en el modelo no es significativa.

14.15 *a*) La interacción es significativa al nivel de 0.05, con un valor *P* de 0.0166.
b) Ambos efectos principales son significativos.

14.17 *a*) *AB*: $f = 3.83$; significativo;
AC: $f = 3.79$; significativo;
BC: $f = 1.31$; no es significativo;
ABC: $f = 1.63$; no es significativo;
b) *A*: $f = 0.54$; no es significativo;
B: $f = 6.85$; significativo;
C: $f = 2.15$; no es significativo;
c) La presencia de la interacción *AC* enmascara el efecto principal *C*.

14.19 *a*) Esfuerzo cortante: $f = 45.96$ con valor $P < 0.0001$;
Recubrimiento: $f = 0.05$ con valor $P = 0.8299$;
Humedad: $f = 2.13$ con valor $P = 0.1257$;
recubrimiento × humedad: $f = 3.41$ con valor $P = 0.0385$;
recubrimiento × esfuerzo cortante: $f = 0.08$ con valor $P = 0.9277$;
humedad × esfuerzo cortante: $f = 3.15$ con valor $P = 0.0192$;
recubrimiento × humedad × esfuerzo cortante: $f = 1.93$ con valor $P = 0.1138$.
b) La mejor combinación parece ser sin recubrimiento, humedad media y nivel de esfuerzo cortante de 20.

14.21

Efecto	*f*	*P*
Temperatura	14.22	< 0.0001
Superficie	6.70	0.0020
HRC	1.67	0.1954
T × S	5.50	0.0006
T × HRC	2.69	0.0369
S × HRC	5.41	0.0007
T × S × HRC	3.02	0.0051

14.23 *a*) Sí; marca × tipo; marca × temperatura;
b) sí;
c) marca *Y*, detergente en polvo, alta temperatura.

14.25 *a*)

Efecto	*f*	*P*
Tiempo	543.53	< 0.0001
Temperatura	209.79	< 0.0001
Solvente	4.97	0.0457
Tiempo × temperatura	2.66	0.1103
Tiempo × solvente	2.04	0.1723
Temperatura × solvente	0.03	0.8558
Tiempo × temperatura × solvente	6.22	0.0140

Aunque las tres interacciones bilaterales son insignificantes podrían estar enmascaradas por la interacción trilateral significativa.

14.27 *a*) $f = 1.49$; no hay interacción significativa;
b) f(operadores) $= 12.45$; significativo;
f(filtros) $= 8.39$; significativo;
c) $\hat{\sigma}_\alpha^2 = 0.1777$ (filtros);
$\hat{\sigma}_\beta^2 = 0.3516$ (operadores);
$s^2 = 0.185$

14.29 *a*) $\hat{\sigma}_\beta^2, \hat{\sigma}_\gamma^2, \hat{\sigma}_{\alpha\gamma}^2$ son significativos;
b) $\hat{\sigma}_\gamma^2$ y $\hat{\sigma}_{\alpha\gamma}^2$ son significativos

14.31 *a*) Modelo combinado;

b) Material: $f = 47.42$ con valor $P < 0.0001$; marca: $f = 1.73$ con valor $P = 0.2875$; material × marca: $f = 16.06$ con valor $P = 0.0004$;
c) no

Capítulo 15

15.1 *B* y *C* son significativos al nivel 0.05

15.3 Los factores *A*, *B* y *C* tienen efectos negativos sobre el compuesto de fósforo y el factor *D* tiene un efecto positivo. Sin embargo, la interpretación del efecto de los factores individuales debería implicar el uso de las gráficas de interacción.

15.5 Efectos significativos:
A: $f = 9.98$; *BC*: $f = 19.03$.
Efectos insignificantes:
B: $f = 0.20$; *C*: $f = 6.54$; *D*: $f = 0.02$; *AB*: $f = 1.83$;
AC: $f = 0.20$; *AD*: $f = 0.57$; *BD*: $f = 1.83$;
CD: $f = 0.02$. Como la interacción *BC* es significativa, se investigaría más sobre *B* y sobre *C*.

15.9 *a*) $b_A = 5.5$, $b_B = -3.25$ y $b_{AB} = 2.5$;
b) Los valores de los coeficientes son de la mitad de los efectos;
c) $t_A = 5.99$ con valor $P = 0.0039$;
$t_B = -3.54$ con valor $P = 0.0241$;
$t_{AB} = 2.72$ con valor $P = 0.0529$;
$t_2 = F$.

15.11 *a*) $A = -0.8750$, $B = 5.8750$, $C = 9.6250$, $AB = -3.3750$, $AC = -9.6250$, $BC = 0.1250$ y $ABC = -1.1250$;
B, C, AB y *AC* parecen importantes con base en sus magnitudes.

b)

Efectos	Valor *P*
A	0.7528
B	0.0600
C	0.0071
AB	0.2440
AC	0.0071
BC	0.9640
ABC	0.6861

c) Sí;
d) A un nivel alto de *A*, *C* esencialmente no tiene efecto. A un nivel bajo de *A*, *C* tiene un efecto positivo.

15.13 *a*)

Máquina 1	2	3	4
(1)	*c*	*a*	*ac*
ab	*d*	*b*	*ad*
cd	*e*	*acd*	*ae*
ce	*abc*	*ace*	*bc*
de	*abd*	*ade*	*bd*
abcd	*abe*	*bcd*	*be*
abce	*cde*	*bce*	*acde*
abde	*abcde*	*bde*	*bcde*

b) *ABD, CDE, ABCDE* (un posible diseño)

15.15 *a*) x_2, x_3, x_1x_2 y x_1x_3;
b) Curvatura: valor $P = 0.0038$;
c) Un punto de diseño adicional diferente de los originales.

15.17 (0, –1), (0, 1), (–1, 0), (1, 0) podría utilizarse.

15.19 *a*) Con *BCD* como el contraste de definición, el bloque principal contiene (1), *a, bc, abc, bd, abd, cd, acd*;

b)

Bloque 1	Bloque 2
(1)	*a*
bc	*abc*
abd	*bd*
acd	*cd*

confundido por *ABC*;

c) El contraste de definición *BCD* produce los siguientes alias: $A \equiv ABCD$, $B \equiv CD$, $C \equiv BD$, $D \equiv BC$, $AB \equiv ACD$, $AC \equiv ABD$ y $AD \equiv ABC$. Puesto que *AD* y *ABC* están confundidos con los bloques sólo hay dos grados de libertad para el error en las interacciones no confundidas.

Fuente de variación	Grado de libertad
A	1
B	1
C	1
D	1
Bloques	1
Error	2
Total	7

15.21 *a*) Con el contraste de definición *ABCE* y *ABDF* el bloque principal contiene (1), *ab, acd, bcd, ce, abce, ade, bde, acf, bcf, df, abdf, aef, bef, cdef, abcdef*;

b) $A \equiv BCE \equiv BDF \equiv ACDEF$,
$AD \equiv BCDE \equiv BF \equiv ACEF$,
$B \equiv ACE \equiv ADF \equiv BCDEF$,
$AE \equiv BC \equiv BDEF \equiv ACDF$,
$C \equiv ABE \equiv ABCDF \equiv DEF$,
$AF \equiv BCEF \equiv BD \equiv ACDE$,
$D \equiv ABCDE \equiv ABF \equiv CEF$,
$CE \equiv AB \equiv ABCDEF \equiv DF$,
$E \equiv ABC \equiv ABDEF \equiv CDF$,
$DE \equiv ABCD \equiv ABEF \equiv CF$,
$F \equiv ABCEF \equiv ABD \equiv CDE$,
$BCD \equiv ADE \equiv ACF \equiv BEF$,
$AB \equiv CE \equiv DF \equiv ABCDEF$,
$BCF \equiv AEF \equiv ACD \equiv BDE$,
$AC \equiv BE \equiv BCDF \equiv ADEF$;

Fuente de variación	Grados de libertad
A	1
B	1
C	1
D	1
E	1
F	1
AB	1
AC	1
AD	1
BC	1
BD	1
CD	1
CF	1
Error	2
Total	15

15.23

Fuente	gl	SC	CM	*f*	*P*
A	1	6.1250	6.1250	5.81	0.0949
B	1	0.6050	0.6050	0.57	0.5036
C	1	4.8050	4.8050	4.56	0.1223
D	1	0.2450	0.2450	0.23	0.6626
Error	3	3.1600	1.0533		
Total	7	14.9400			

15.25

Fuente	gl	SC	CM	*f*	*P*
A	1	388,129.00	388,129.00	3585.49	0.0001
B	1	277,202.25	277,202.25	2560.76	0.0001
C	1	4692.25	4692.25	43.35	0.0006
D	1	9702.25	9702.25	89.63	0.0001
E	1	1806.25	1806.25	16.69	0.0065
AD	1	1406.25	1406.25	12.99	0.0113
AE	1	462.25	462.25	4.27	0.0843
BD	1	1156.00	1156.00	10.68	0.0171
BE	1	961.00	961.00	8.88	0.0247
Error	6	649.50	108.25		
Total	15	686,167.00			

Todos los efectos principales son significativos al nivel 0.05; *AD*, *BD* y *BE* son también significativos al nivel 0.05.

15.27 El bloque principal contiene *af, be, cd, abd, ace, bcf, def, abcdef.*

15.29 $A \equiv BD \equiv CE \equiv CDF \equiv BEF \equiv ABCF \equiv ADEF \equiv ABCDE$;
$B \equiv AD \equiv CF \equiv CDE \equiv AEF \equiv ABCE \equiv BDEF \equiv ABCDF$;
$C \equiv AE \equiv BF \equiv BDE \equiv ADF \equiv CDEF \equiv ABCD \equiv ABCEF$;
$D \equiv AB \equiv EF \equiv BCE \equiv ACF \equiv BCDF \equiv ACDE \equiv ABDEF$;
$E \equiv AC \equiv DF \equiv ABF \equiv BCD \equiv ABDE \equiv BCEF \equiv ACDEF$;
$F \equiv BC \equiv DE \equiv ACD \equiv ABE \equiv ACEF \equiv ABDF \equiv BCDEF$.

15.31 $x_1 = 1$ y $x_2 = 1$

15.33 *a*) Sí;
b) i) $E(\hat{y}) = 79.00 + 5.281A$;
ii) $\text{Var}(\hat{y}) = 6.22^2\,\sigma_Z^2 + 5.70^2A^2\,\sigma_Z^2 + 2(6.22)(5.70)A\,\sigma_Z^2$;
c) velocidad a bajo nivel;
d) velocidad a bajo nivel;
e) sí

15.35 $\hat{y} = 12.7519 + 4.7194x_1 + 0.8656x_2 - 1.4156x_3$; las unidades están centradas y a escala; prueba de falta de ajuste, $F = 81.58$, con valor $P < 0.0001$.

15.37 *AFG, BEG, CDG, DEF, CEFG, BDFG, BCDE, ADEG, ACDF, ABEF* y *ABCDEFG.*

Capítulo 16

16.1 $x = 7$ con valor $P = 0.1719$; no rechace H_0.

16.3 $x = 3$ con valor $P = 0.0244$; rechace H_0.

16.5 $x = 4$ con valor $P = 0.3770$; no rechace H_0.

16.7 $x = 4$ con valor $P = 0.1335$; no rechace H_0.

16.9 $w = 43$; no rechace H_0.

16.11 $w_+ = 17.5$; no rechace H_0.

16.13 $w_+ = 15$ con $n = 13$; rechace H_0 a favor de $\mu_1 - \mu_2 < 8$.

16.15 $u_1 = 4$; la afirmación no es válida

16.17 $u_2 = 5$; *A* opera durante más tiempo.

16.19 $u = 15$; no rechace H_0.

16.21 $h = 10.58$; los tiempos de operación son diferentes.

16.23 $v = 7$ con valor $P = 0.910$; muestra aleatoria.

16.25 $v = 6$ con valor $P = 0.044$; no rechace H_0.

16.27 $v = 4$; muestra aleatoria.

16.29 0.70

16.31 0.995

16.33 *a*) $r_s = 0.39$; *b*) no rechace H_0.

16.35 *a*) $r_s = 0.72$; *b*) rechace H_0, de manera que $\rho > 0$.

16.37 *a*) $r_s = 0.71$; *b*) rechace H_0, de manera que $\rho > 0$.

Capítulo 18

18.1 $p^* = 0.173$

18.3 *a*) $\pi(p \mid x = 1) = 40p(1 - p)^3/0.2844$; $0.05 < p < 0.15$:
b) $p^* = 0.106$

18.5 *a*) *beta*(95, 45); *b*) 1

18.7 $8.077 < \mu < 8.692$

18.9 *a*) 0.2509; *b*) $68.71 < \mu < 71.69$; *c*) 0.0174

18.13 $p^* = \frac{6}{x + 2}$

18.15 2.21

Índice analítico

A

Análisis de varianza (ANOVA), 254, 507
- de dos factores, 565
- de tres factores, 579
- de un factor, 509
  - comparación de, 520
  - contraste de, 520
  - de un solo grado de libertad, 520
  - efecto del tratamiento, 510
  - media grande, 510
  - suma de cuadrados de los contrastes, 521
  - tratamiento, 509
- tabla de, 415

Aplicaciones bayesianas, 710
Aproximación
- de binomial a hipergeométrica, 155
- de grados de libertad de Satterthwaite, 289
- de normal a binomial, 187, 188
- de Poisson a binomial, 163

B

Bernoulli
- ensayo de, 144
- proceso de, 144
- variable aleatoria, 83

Bloques, 509

C

Cadena Markov de Monte Carlo, 710
Coeficiente
- de confianza, 269
- de correlación, 125, 431
  - de la población, 432
  - de rangos, 675
  - muestral, 432
  - producto-momento de Pearson, 432
- de determinación, 407, 433, 462
  - ajustado, 464
- de variación, 471

Combinación, 50
Complemento de un evento, 39
Confianza
- coeficiente de, 269
- grado de, 269
- límites, 269, 271

Contrastes ortogonales, 522
Control de calidad, 681
- dentro de control, 682
- fuera de control, 682
- gráfica, 681, 682
- límites del, 683

Corrección de continuidad, 190
Covarianza, 119, 123
Cuadrado(s)
- medio(s), 415
  - del error, 284
  - esperados, 548

Cuantiles, 255
Curva característica de operación, 335

D

Datos históricos, 30
Desviación, 120
- estándar, 120, 122, 135
  - muestral, 15, 16

Diagrama(s)
- de árbol, 36
- de dispersión, 3
- de Venn, 40

Diseño
- central compuesto, 640
- completamente aleatorizado, 8, 509
- de bloques completos aleatorizado, 533

de experimento
central compuesto, 640
completamente aleatorizado, 532
contraste en el, 599
de bloques, 532
de bloques aleatorizados, 533
factor de ruido, 644
factores de control, 644
factorial fraccionario, 598, 612, 626, 627
ortogonal, 617
relación de definición, 627
resolución, 637
Distribución, 23
beta, 201
binomial, 104, 143-145, 153, 155, 175, 188
negativa, 143, 158-60
media de la, 147
varianza de la, 147
chi cuadrada, 200
condicional, 99
conjunta, 103
continua
beta, 201
chi cuadrada, 200
de Weibull, 203, 204
exponencial, 195
gamma, 195
logarítmica normal, 201
normal, 172
uniforme, 171
de Erlang, 206-207
de muestreo, 232
de la media, 233
de Poisson, 143, 161, 162
media de la, 162
varianza de la, 162
de probabilidad, 84
condicional, 99
conjunta, 94, 95, 102
continua, 87
discreta, 84
marginal, 97
media de la, 111
varianza de la, 119
de razón de varianza, 253
de Weibull, 203
función de distribución acumulativa para, 204
media de la, 203
tasa de fallas de, 204, 205
varianza de la, 203
discreta
binomial, 143, 144, 158, 159
de Poisson, 161, 162
geométrica, 158, 160
hipergeométrica, 152, 153
multinomial, 143, 149
empírica, 254
exponencial, 104, 194, 195
media de la, 196
negativa, 196
propiedad de falta de memoria de, 197
relación con el proceso de Poisson, 196
varianza de la, 196
F, 251-254
gamma, 194-195
media de la, 196
relación con el proceso de Poisson, 196
varianza de la, 196
gaussiana, 19, 172
geométrica, 143, 158, 160
media de la, 160
varianza de la, 160
hipergeométrica, 152-154, 175
media de la, 154
multivariada, 156
varianza de la, 154
hipergeométrica multivariada, 156
logarítmica normal, 201
media de la, 202
varianza de la, 202
marginal, 97, 101, 102
conjunta, 103
multinomial, 143, 149
normal, 19, 172, 173, 188
bivariada, 431
curva normal, 172-175
desviación estándar de la, 175
estándar, 177
media de la, 175
varianza de la, 175
posterior, 711
previa, 710

rectangular, 171
sesgada, 23
simétrica, 23
t, 246-250
uniforme, 171
continua, 171

E

Ecuaciones normales para la regresión lineal, 444
Efecto de enmascaramiento, 563
Eliminación hacia atrás, 479
Error(es)
en la estimación de la media, 272
estándar de la media, 277
experimental, 509
suma de cuadrados del, 402
tipo I, 322
tipo II, 323
Espacio muestral, 35
continuo, 83
discreto, 83
partición del, 57
Esperanza matemática, 111, 112, 115
Estadística
descriptiva, 3, 9
inferencial, 1
Estadístico, 228
C_p, 491
de prueba, 322
Estimación, 12, 142, 266
de dos proporciones, 300
de la diferencia de dos medias muestrales, 285
de la probabilidad máxima, 307, 308, 312
de la proporción de varianzas, 305
de observaciones en pares, 291
de proporciones, 296
de una sola varianza, 303
Estimación de máxima verosimilitud, 307, 308, 710
residual, 550
restringida, 550
Estimado(s), 12
agrupado de la varianza, 287
bayesianos, 717
bajo la pérdida de error absoluto, 718
bajo la pérdida del cuadrado del error, 717
de una sola media, 269
del intervalo, 268
bayesiano, 715
puntual, 266, 268
error estándar, 276
Estimador, 266
de probabilidad máxima, 308-310
eficiente, 267
insesgado, 266, 267
método de momentos, 314, 315
puntual, 266, 268
Estudio
observacional, 3, 29
retrospectivo, 30
Evento(s), 38
mutuamente excluyentes, 40
Experimento
binomial negativo, 158
de efectos aleatorios
componentes de la varianza, 549
de efectos fijos, 547
de Poisson, 161
factorial, 561
ANOVA de los factores, 565
ANOVA de tres factores, 579
cuadrados medios agrupados, 583
efectos aleatorios, 589
efectos de enmascaramiento, 563
efectos principales, 562
en bloques, 583
factor, 507
interacción, 562
modelo mixto, 591
nivel, 507
tratamiento, 507
factorial 2^k, 597
ajuste de regresión, 612
alias, 628
corridas centrales, 620
diseño ortogonal, 617
diseños de Plackett-Burman, 638
factorial fraccionario, 626
filtrado de factores, 598
generación del diseño, 627
gráficas de diagnóstico, 604

relación definitoria, 627
resolución, 637

F

Factor, 28, 507
Factorial, 47
Falta de ajuste, 418
Frecuencia relativa, 22, 31, 111
Función(es)
de densidad de probabilidad, 88, 89
conjunta, 96
de distribución acumulativa, 85, 90
de masa de probabilidad, 84
conjunta, 95
de pérdida
del cuadrado del error, 717
del error absoluto, 718
de probabilidad, 84, 308
gamma, 194
incompleta, 199
generadoras de momentos, 218

G

Grados de libertad, 15, 16, 200, 244, 246
aproximación de Satterthwaite de, 289
Gráfica(s)
de caja, 3, 24, 25
de control
de atributos, 697
de variables, 684
gráfica cusum, 705
p, 697
R, 688
S, 695
U, 704
$\bar{X}$, 686
de cuantiles, 254, 255
cuantiles normales, 256, 257
de probabilidad, 254
normal, 254
de puntos, 3, 8, 32
de tallo y hojas, 3, 21, 22, 31
p, 697
R, 688
S, 695
U, 704
$\bar{X}$, 686
función característica de operación, 691

H

Hipótesis, 320
alternativa, 320
estadística, 319
nula, 320
prueba de, 320, 321
Histograma(s), 22
de probabilidad, 86

I

Independencia, 62, 65, 67, 68
estadística, 101-103
Inferencia
bayesiana, 710
estadística, 3, 225, 265
Interacción, 28, 562
Intersección de eventos, 39
Intervalo
bayesiano, 715
posterior, 317
de confianza, 269, 270, 281, 317
de una muestra grande, 276
interpretación de, 289
para el cociente de las desviaciones estándar, 306
para el cociente de las varianzas, 306
para la desviación estándar, 304
para la diferencia de dos medias, 285-288, 290
para la diferencia de dos proporciones, 300, 301
para observaciones en pares, 293
para una sola media, 269-272, 275
unilateral, 273
para una sola proporción, 297
para una sola varianza, 304
de predicción, 277, 278, 281
para una observación futura, 278, 279
unilateral, 279
de tolerancia, 280, 281
posterior bayesiano, 317

J

Jacobiano, de la transformación, 213

L

Límite(s)
de confianza unilateral, 273
de tolerancia, 280
del método no paramétrico, 674
unilaterales, 281

M

Media, 19, 111, 112, 114, 115
muestral, 111
poblacional, 12, 16
recortada, 12
Método(s)
de distribución libre, 655
de la regla, 37
de mínimos cuadrados, 394, 396
no paramétricos, 655
límites de tolerancia, 674
prueba de Kruskall-Wallis, 668
prueba de la suma de rangos de Wilcoxon, 665
prueba de rachas, 671
prueba de rango con signo, 660
prueba del signo, 656
Metodología bayesiana, 265, 709
Metodología de respuesta superficial, 447, 639, 640
factor(es)
de control, 644
de ruido, 644
modelo de segundo orden, 640
Moda, 713
distribución normal, 174
Modelo
de efectos aleatorios, 547, 548
lineal, 133
Momentos, 218
Muestra, 1, 2, 225, 226
aleatoria, 227
simple, 7
desviación estándar de la, 3, 15, 16, 30, 31, 229, 230
media de la, 3, 11, 12, 19, 30-32, 225, 228
mediana de la, 3, 11, 12, 30, 31, 228
moda de la, 228
rango de la, 15, 30, 31, 229
sesgada, 7
tamaño de la, 7
varianza de la, 15, 16, 30, 225, 229
Muestreo
aleatorio, 225
de aceptación, 153
Multicolinealidad, 476

N

Nivel
de calidad aceptable, 705
de calidad rechazable, 705
de significancia, 323, 332

O

Observaciones en pares, 291

P

Parámetro(s), 12, 142
de distribución, 104
de la población, 16, 104
Permutación, 47
circular, 49
Perspectiva
bayesiana, 710
condicional, 710
Población, 2, 4, 225, 226
media de la, 226
parámetro de la, 16, 104
tamaño de la, 226
varianza de la, 226
Potencia de una prueba, 329
Predictor lineal, 498
Probabilidad, 35, 52, 53
condicional, 62-66, 68, 75, 76
de cobertura, 715
de un evento, 52
frecuencia relativa, 55, 709
función de masa, 84
indiferencia, 55, 709
método subjetivo, 709, 710
regla aditiva, 56

subjetiva, 709, 710
total, 72, 73
Proceso de Poisson, 161, 196
relación con la distribución gamma, 196
Promedio, 111
Prueba(s)
de Bartlett, 516
de bondad de ajuste, 210, 255, 317, 370, 371
de Cochran, 518
de comparación múltiple, 523
de Duncan, 527
de Dunnett, 528
de Tukey, 526
tasa de error por experimento-familia, 525
de Duncan de rango múltiple, 527
de Dunnett, 528
de hipótesis, 19, 266, 319
bondad de ajuste, 210, 255, 370, 371
de dos colas, 330
de una cola, 330
dos medias con varianza desconocidas pero iguales, 343
dos medias con varianzas desconocidas y desiguales, 345
dos varianzas, 366
estadísticos de prueba, 326
F parcial, 466
observaciones en pares, 345
propiedades importantes, 329
prueba de homogeneidad, 376
prueba de independencia, 373
prueba de varias proporciones, 377
región crítica, 322
selección del tamaño de la muestra, 349, 352
sobre dos medias, 342
tamaño de prueba, 323
una sola muestra, 336
una sola muestra, varianza conocida, 336
una sola muestra, varianza desconocida, 340
una sola proporción, 360
una sola varianza, 366
valor crítico, 322
valor *P*, 331, 333
de Kruskall-Wallis, 668
de la suma de rangos de Wilcoxon, 665
de rachas, 671
de rango con signo, 660
de Tukey, 526
del signo, 656
F parcial, 466
para la igualdad de varianzas, 516
de Bartlett, 516
de Cochran, 518
sobre una sola proporción, 360
Puntos de inflexión, distribución normal, 174

R

R^2, 407, 462
ajustada, 464
Rango intercuartilar, 24, 25
Regla
aditiva, 56
de Bayes, 72, 75
de eliminación, 73-75
de la multiplicación, 44
del producto, 65
multiplicativa, 65
Regresión, 20
Regresión lineal
a través del origen, 413
ANOVA, 414
coeficiente de determinación, 407
coeficiente de regresión, 392
correlación, 430
cuadrados medios, 394
ecuación normal, 396
error aleatorio, 391
error experimental puro, 419
falta de ajuste de la, 418
intervalo de predicción, 410, 411
línea ajustada, 392
modelo
empírico, 391
estadístico, 391
múltiple, 390, 443
ANOVA, 455
ecuaciones normales, 444
inferencia, 455
matriz de varianza-covarianza, 453
matriz TESTADA, 483
multicolinealidad, 476

polinomial, 446
R^2 ajustada, 464
residuales estudentizados, 483
residuales R de Student, 483
selección de variables, 456
suma de cuadrados de regresión, 460
suma de cuadrados del error, 460
valor extremo, 484
variables ortogonales, 467
predicción, 408
prueba de linealidad, 416
regresor, 389
residual, 395
respuesta media, 394, 409
selección del modelo, 476, 487
simple, 389, 390
sobreajuste, 408
suma
de cuadrados de la regresión, 461
de cuadrados del error, 415
de cuadrados total, 414
transformación de datos, 424
valor ajustado, 416
variable
categórica, 472
dependiente, 389
independiente, 389
Regresión logística, 497
dosis eficaz, 500
razón de probabilidad, 500
Regresión no lineal, 496
datos de conteo, 497
logística, 497
respuesta binaria, 497
Regresión polinomial, 443, 446
Regresión por etapas, 479
Residual, 395, 427

S

Selección
del modelo, 476
eliminación hacia atrás, 480
estadístico C_p, 491
métodos secuenciales, 476
PRESS, 487, 488
regresión por etapas, 480
selección hacia adelante, 479
hacia adelante, 479
Sesgo, 227
Suma de cuadrados
de predicción, 487, 488
del error, 402, 415
falta de ajuste, 419
identidad, 510, 536, 567
regresión, 415
total, 407
tratamiento, 511, 522, 536
Superficie de respuesta, 642, 648
diseño de parámetro robusto, 644

T

Tabla de contingencia, 373
frecuencia marginal, 374
Tamaño de la muestra, 7
en la estimación
de una media, 272
de una proporción, 298
en la prueba de hipótesis, 351
Tasa
de error por experimento-familia, 525
de fallas, 204, 205
Teorema
de Chebyshev, 135-137, 148, 155, 180, 186
del límite central, 233, 234, 238
Transformación de variables
continuas, 213, 214
discretas, 212
Tratamiento
efecto negativo del, 563
efecto positivo del, 563

U

Unidad experimental, 9, 286, 292, 562
Unión de eventos, 40

V

Validación cruzada, 487
Valor(es)
esperado, 112-115
extremo, 24, 279, 484

P, 4, 109, 331-333
Variabilidad, 8, 9, 14-16, 119, 135, 228, 251, 253
entre/dentro de muestras, 253, 254
Variable
aleatoria, 81
binomial, 144, 147, 158
continua, 84
chi cuadrada, 244
de Bernoulli, 83, 147
de Poisson, 161, 162
discreta, 83, 84
función no lineal de la, 133
hipergeométrica, 143, 153
media de la, 111, 114
multinomial, 149
normal, 173
transformación, 211
uniforme continua, 171
uniforme discreta, 150
varianza de la, 119, 122
categórica, 472
ficticia, 472
indicadora, 472
ortogonales, 467
Varianza, 119, 120, 122
muestral, 16
agrupada, 287
poblacional, 16

Al elaborar la novena edición de esta obra, el interés principal de los autores no fue tan sólo incluir material nuevo, sino brindar claridad y una mejor exposición, así como conservar el equilibrio entre la teoría y las aplicaciones.

Con la finalidad de motivar al estudiante, muchos ejercicios se refieren a aplicaciones científicas y de ingeniería en la vida real. En varios capítulos se agregaron proyectos para la clase y más estudios de caso, con el fin de ayudar a los usuarios a entender los métodos estadísticos que se presentan en el contexto de una situación cotidiana. Para lograr que los estudiantes adquieran experiencia en la lectura e interpretación de listas de resultados y gráficas por computadora, los estudios de caso muestran impresiones de listas de resultados por computadora y material gráfico generado con los programas SAS y MINITAB. En algunas situaciones, los ejemplos y los estudios de caso se complementan con diversos tipos de gráficas residuales, cuantilares, de probabilidad normal y de otros tipos.

Todos los conjuntos de datos asociados con los ejercicios
están disponibles para descargar del sitio Web
http://www.pearsonenespañol.com/walpole

ISBN 978 607-32-1417-9
90000
9 786073 214179

Visítenos en:
www.pearsonenespañol.com